IRE Transactions on INFORMATION THEORY

Vol. IT-2, No. 3 September, 1956

1956 SYMPOSIUM ON INFORMATION THEORY

held at

Massachusetts Institute of Technology
Cambridge, Massachusetts

September 10-12, 1956

PUBLISHED BY THE
Professional Group on Information Theory

IRE PROFESSIONAL GROUP ON INFORMATION THEORY

The Professional Group on Information Theory is an organization, within the framework of the IRE, of members with principal professional interest in Information Theory. All members of the IRE are eligible for membership in the Group and will receive all Group publications upon payment of prescribed assessments.

Annual Assessment: $2.00

IRE TRANSACTIONS®
on Information Theory

Published by the Institute of Radio Engineers, Inc., for the Professional Group on Information Theory, 1 East 79th Street, New York 21, N. Y. Responsibility for the contents rests upon the authors, and not upon the IRE, the Group or its members. Individual copies available for sale to IRE-PGIT members at $3.00, to IRE members at $4.50 and to nonmembers at $9.00.

TRANSACTIONS
of the
1956 SYMPOSIUM ON INFORMATION THEORY

held at

Massachusetts Institute of Technology, Cambridge, Massachusetts

September 10-12, 1956

Organized by
The Professional Group on Information Theory, Institute of Radio Engineers

In cooperation with
The Research Laboratory of Electronics, Massachusetts Institute of Technology

and sponsored by

The International Scientific Radio Union (URSI)
The Office of Naval Research
The Signal Corps Engineering Laboratories
The Air Research and Development Command

These papers are to be presented at the 1956 Symposium on Information Theory. They are published prior to the Symposium to allow informed discussion at the meeting.

CONTENTS AND ABSTRACTS

Page

CODING I

"The Zero Error Capacity of a Noisy Channel," by C. E. Shannon..... 8

The zero error capacity C_0 of a noisy channel is defined as the least upper bound of rates at which it is possible to transmit information with zero probability of error. Various properties of C_0 are studied; upper and lower bounds and methods of evaluation of C_0 are given. Inequalities are obtained for the C_0 relating to the "sum" and "product" of two given channels. The analogous problem of zero error capacity C_{0F} for a channel with a feedback link is considered. It is shown that while the ordinary capacity of a memoryless channel with feedback is equal to that of the same channel without feedback, the zero error capacity may be greater. A solution is given to the problem of evaluating C_{0F}.

"A Linear Circuit Viewpoint on Error-Correcting Codes," by D. A. Huffman.. 20

A linear binary filter has as its output a binary sequence, each digit of which is the result of a parity check on a selection of preceding output digits and of present and preceding digits of the filter input sequence. The terminal properties of these filters may be described by transfer ratios of polynomials in a delay operator. If two binary filters have transfer ratios which are reciprocally related then the filters are mutually inverse in the sense that, in a cascade connection, the second filter unscrambles the scrambling produced by the first. The coding of a finite sequence of binary information digits for protection against noise may be accomplished by a binary sequence filter, the output of which becomes the sequence to be transmitted. The inverse filter is utilized at the receiver.

CODING II

"Theory of Information Feedback Systems," by S. S. L. Chang..... 29

A general information feedback system is defined and formulated in a way broad enough to allow coded or uncoded channels with total or partial information feedback. Basic theorems governing change in information rate and reliability are derived with full consideration of the transition probabilities of both direct and feedback channels, including message words as well as the confirmation--denial signal.

"A Linear Coding for Transmitting a Set of Correlated Signals," by H. P. Kramer and M. V. Mathews..... 41

A coding scheme is described for the transmission of n continuous correlated signals over m channels, m being equal to or less than n. Each of the m signals is a linear combination of the n original signals.

"On an Application of Semi-Group Methods to Some Problems in Coding," by M. P. Schutzenberger..... 47

We give an abstract model of some sort of language and try to show how semi-group concepts apply fruitfully to it with the hope that some of them may be of interest to specialists working on natural languages. In a first part, the model and its main properties are discussed at a concrete level on the simplest cases: coding and decoding with length-bounded codes. In a second part a selection of theorems are proved whenever the necessary semi-group-theoretic preliminaries are not exacting.

AUTOMATA

"The Logic Theory Machine," by A. Newell and H. A. Simon..... 61

In this paper we describe a complex information processing system, which we call the logic theory machine, that is capable of discovering proofs for theorems in symbolic logic. This system relies heavily on heuristic methods similar to those that have been observed in human problem solving activity. The present paper is concerned with specification of the system, and not with its realization in a computer.

"Tests on a Cell Assembly Theory of the Action of the Brain, Using a Large Digital Computer," by N. Rochester, J. H. Holland, L. H. Haibt and W. L. Duda..... 80

Theories by D. O. Hebb and P. M. Milner on how the brain works were tested by simulating neuron nets on the IBM Type 704 Electronic Calculator. The cell assemblies do not yet act just as the theory requires, but changes in the theory and the simulation offer promise for further experimentation.

INFORMATION SOURCES

"The Measurement of Third Order Probability Distributions of Television Signals," by W. F. Schreiber..... 94

A device has been built for the rapid, automatic measurement of the third order probability density of video signals. Examples are presented of second and third order distributions, and of entropies calculated for a variety of scenes.

"Gap Analysis and Syntax," by V. H. Yngve..... 106

A statistical procedure has been tried as a method of investigating the structure of language with the aid of data processing machines. The frequency of gaps of various lengths between occurrences of two specified words is counted. The results are compared with what would be expected if the occurrences of the two words were statistically independent. Deviations from the expected number give clues to the constraints that operate between words in a language.

"Three Models for the Description of Language," by A. N. Chomsky..... 113

We investigate several conceptions of linguistic structure to determine whether or not they can provide simple and "revealing" grammars that generate all of the sentences of English and only these. We find that no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar. We formalize the notion of "phrase structure" and show that this gives us a method for describing language which is essentially more powerful. We study the properties of a set of grammatical transformations, showing that the grammar of English is materially simplified if phrase-structure description is limited to a kernel of simple sentences from which all other sentences are constructed by repeated transformations, and that this view of linguistic structure gives a certain insight into the use and understanding of language.

INFORMATION USERS

"Some Studies in the Speed of Visual Perception," by G. C. Sziklai..... 125

Statistical studies of television signals indicated a high degree of correlation between successive elements, lines and frames. Some tests were devised to measure the perception speed of observers. These tests included certain reading and character recognition tests and finally a test consisting of object recognition in precisely measured periods was devised. Several series of these tests indicated that the visual perception speed of a normal observer is between 30 and 50 bits per second, that this value holds for periods of one-tenth to two seconds, and that the first thing observed is the center of the picture.

"Human Memory and the Storage of Information," by G. A. Miller..... 129

The amount of selective information in a message can be increased either by increasing the variety of the symbols from which it is composed or by increasing the length of the message. The variety of the symbols is far less important than the length of the message in controlling what human subjects are able to remember.

"The Human Use of Information III. Decision-Making in Signal Detection and Recognition Situations Involving Multiple Alternatives," by J. A. Swets and T. G. Birdsall..... 138

A general theory of signal detectability, constructed after the model provided by decision theory, is applied to the performance of the human observer faced with the problem of choosing among multiple signal alternatives on the basis of a fixed, finite observation interval. The results indicate that a highly simplified theory is adequate for prediction of the obtained payoff and response-frequency tables to within a few per cent. They also indicate the fairly large extent to which intelligence may influence a sensory process usually assumed to involve fixed parameters.

OPTIMUM MEAN-SQUARE OPERATIONS

"On Optimum Non-Linear Extraction and Coding Filters," by A. V. Balakrishnan and R. F. Drenick..... 166

The problem of determining optimal non-linear least-square filters is solved for a class of stationary time series. This theory is then used as the basis for developing a band-width reduction scheme using non-linear encoding and decoding filters, for the same class of signals. A simple illustrative example is included.

"Final-Value Systems with Gaussian Inputs," by R. C. Booton, Jr..... 173

A final-value system controls a response variable r(t) over a time interval (0,T) with the objective of minimizing the difference between a desired value ρ and the final response value r(T). Physical limitations of the element being controlled result in a maximum-value constraint on the system velocity r'(t). Earlier results suggest that a system consisting of an estimator followed by a "bang-bang" servo is approximately optimum. The estimator uses the input to produce an estimate ρ^* of the desired response and the servo results in a system velocity as large in magnitude as possible and with the same sign as the difference $\rho^* - r$. The present paper shows that this system is the true optimum when the joint distribution of the input and the desired response is Gaussian and the error criterion is minimization of the average of a nondecreasing function of the magnitude of the error.

"An Extension of the Minimum Mean Square Prediction Error Theory for Sampled Data," by M. Blum..... 176

A method is developed for finding the ordinates of a digital filter which will produce a general linear operator of the signal S(t) such that the mean square error of prediction will be a minimum. The input to the filter is sampled at intervals t. The samples contain stationary noise N(j t), a stationary signal component, M(j t), and a nonrandom signal component. The solution is obtained as a matrix equation which relates the ordinates of the digital filter to the autocorrelation properties of M(t) and N(t) and the nature of the prediction operation.

APPLICATIONS

"A New Interpretation of Information Rate," by J. L. Kelly..... 185

If the input symbols to a communication channel represent the outcomes of a chance event on which bets are available at odds consistent with their probabilities (i.e., "fair" odds), a gambler can use the knowledge given him by the received symbols to cause his money to grow exponentially. The maximum exponential rate of growth of the gambler's capital is equal to the rate of transmission of information over the channel. Thus we find a situation in which the transmission rate has significance even though no coding is contemplated.

"An Outline of a Purely Phenomenological Theory of Statistical Thermodynamics: I. Canonical Ensembles," by B. Mandelbrot..... 190

Since the kinetic foundations of thermodynamics are not sufficient in the absence of further hypotheses of randomness, are they necessary in the presence of such hypotheses? The aim of the paper is to show (partly after Szilard) that a substantial part of the results, usually obtained through kinetic arguments, could be obtained by postulating from the outset a statistical distribution for the properties of a system, and following up with a purely phenomenological argument. It is of interest to the communication engineer to have a unified treatment of the foundations of fluctuation phenomena and of methods of fighting noise.

"A Radar Detection Philosophy," by W. McC. Siebert..... 204

This paper attempts to present a short, unified discussion of the radar detection, parameter estimations, and multiple-signal resolution problems--mostly from a philosophical rather than a detailed mathematical point of view. The purpose is to make it possible in at least some limited sense to reason back from appropriate measures of desired radar performance to specifications of the necessary values of the related radar parameters.

THE ZERO ERROR CAPACITY OF A NOISY CHANNEL

Claude E. Shannon
Bell Telephone Laboratories, Murray Hill, New Jersey
Massachusetts Institute of Technology, Cambridge, Mass.

Abstract

The zero error capacity C_0 of a noisy channel is defined as the least upper bound of rates at which it is possible to transmit information with zero probability of error. Various properties of C_0 are studied; upper and lower bounds and methods of evaluation of C_0 are given. Inequalities are obtained for the C_0 relating to the "sum" and "product" of two given channels. The analogous problem of zero error capacity C_{0F} for a channel with a feedback link is considered. It is shown that while the ordinary capacity of a memoryless channel with feedback is equal to that of the same channel without feedback, the zero error capacity may be greater. A solution is given to the problem of evaluating C_{0F}.

Introduction

The ordinary capacity C of a noisy channel may be thought of as follows. There exists a sequence of codes for the channel of increasing block length such that the input rate of transmission approaches C and the probability of error in decoding at the receiving point approaches zero. Furthermore, this is not true for any value higher than C. In some situations it may be of interest to consider, rather than codes with probability of error approaching zero, codes for which the probability is zero and to investigate the highest possible rate of transmission (or the least upper bound of these rates) for such codes. This rate, C_0, is the main object of investigation of the present paper. It is interesting that while C_0 would appear to be a simpler property of a channel than C, it is in fact more difficult to calculate and leads to a number of as yet unsolved problems.

We shall consider only finite discrete memoryless channels. Such a channel is specified by a finite transition matrix $\|p_i(j)\|$ where $p_i(j)$ is the probability of input letter i being received as output letter j ($i = 1,2,\ldots,a$; $j = 1,2,\ldots,b$) and $\sum_j p_i(j) = 1$. Equivalently, such a channel may be represented by a line diagram such as Fig. 1.

The channel being memoryless means that successive operations are independent. If the input letters i and j are used, the probability of output letters k and l will be $p_i(k)\,p_j(l)$. A sequence of input letters will be called an input word, a sequence of output letters an output word. A mapping of M messages (which we may take to be the integers $1,2,\ldots,M$) into a subset of input words of length n will be called a block code of length n. $R = \frac{1}{n} \log M$ will be called the input rate for this code. Unless otherwise specified, a code will mean such a block code. We will, throughout, use natural logarithms and natural (rather than binary) units of information, since this simplifies the analytical processes that will be employed.

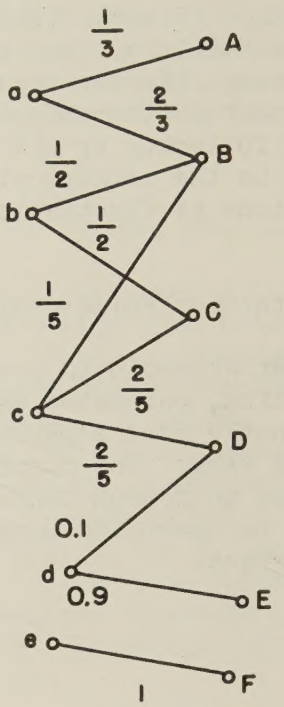


Fig. 1

A decoding system for a block code of length n is a method of associating a unique input message (integer from 1 to M) with each possible output word of length n, that is, a function from output words of length n to the integers 1 to M. The probability of error for a code is the probability when the M input messages are used each with probability 1/M that the noise and the decoding system will lead to an input message different from the one that actually occurred.

If we have two given channels, it is possible to form a single channel from them in two natural ways which we call the sum and product of the two channels. The sum of two

channels is the channel formed by using inputs from either of the two given channels with the same transition probabilities to the set of output letters consisting of the logical sum of the two output alphabets. Thus the sum channel is defined by a transition matrix formed by placing the matrix of one channel below and to the right of that for the other channel and filling the remaining two rectangles with zeros. If $p_i(j)$ and $\| p_i'(j) \|$ are the individual matrices, the sum has the following matrix:

$$\begin{array}{cccccc} p_1(1) & \cdots & p_1(r) & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ p_t(1) & \cdots & p_t(r) & 0 & \cdots & 0 \\ 0 & \cdots & 0 & p_1'(1) & \cdots & p_1'(r') \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & p_t'(1) & \cdots & p_t'(r') \end{array}$$

The product of two channels is the channel whose input alphabet consists of all ordered pairs (i,i') where i is a letter from the first channel alphabet and i' from the second, whose output alphabet is the similar set of ordered pairs of letters from the two individual output alphabets and whose transition probability from (i,i') to (j,j') is $p_i(j)\, p_i'(j')$.

The sum of channels corresponds physically to a situation where either of two channels may be used (but not both), a new choice being made for each transmitted letter. The product channel corresponds to a situation where both channels are used each unit of time. It is interesting to note that multiplication and addition of channels are both associative and commutative, and that the product distributes over a sum. Thus one can develop a kind of algebra for channels in which it is possible to write, for example, a polynomial $\sum a_n K^n$, where the a_n are non-negative integers a and K is a channel. We shall not, however, investigate here the algebraic properties of this system.

The Zero Error Capacity

In a discrete channel we will say that two input letters are adjacent if there is an output letter which can be caused by either of these two. Thus, i and j are adjacent if there exists a t such that both $p_i(t)$ and $p_j(t)$ do not vanish. In Fig. 1, a and c are adjacent, while a and d are not.

If all input letters are adjacent to each other, any code with more than one word has a probability of error at the receiving point greater than zero. In fact, the probability of error in decoding words satisfies

$$P_e \geqslant \frac{M-1}{M}\, p_{min}^n$$

where p_{min} is the smallest (non-vanishing) among the $p_i(j)$, n is the length of the code and M is the number of words in the code. To prove this, note that any two words have a possible output word in common, namely the word consisting of the sequence of common output letters when the two input words are compared letter by letter. Each of the two input words has a probability at least p_{min}^n of producing this common output word. In using the code, the two particular input words will each occur $\frac{1}{M}$ of the time and will cause the common output $\frac{1}{M}\, p_{min}^n$ of the time. This output can be decoded in only one way. Hence at least one of these situations leads to an error. This error, $\frac{1}{M}\, p_{min}^n$, is assigned to this code word, and from the remaining M - 1 code words another pair is chosen. A source of error to the amount $\frac{1}{M}\, p_{min}^n$ is assigned in similar fashion to one of these, and this is a disjoint event. Continuing in this manner, we obtain a total of at least $\frac{M-1}{M}\, p_{min}^n$ as probability of error.

If it is not true that the input letters are all adjacent to each other, it is possible to transmit at a positive rate with zero probability of error. The least upper bound of all rates which can be achieved with zero probability of error will be called the zero error capacity of the channel and denoted by C_0. If we let $M_0(n)$ be the largest number of words in a code of length n, no two of which are adjacent, then C_0 is the least upper bound of the numbers $\frac{1}{n} \log M_0(n)$ when n varies through all positive integers.

One might expect that C_0 would be equal to $\log M_0(1)$, that is, that if we choose the largest possible set of non-adjacent letters and form all sequences of these of length n, then this would be the best error free code of length n. This is not, in general, true, although it holds in many cases, particularly when the number of input letters is small. The first failure occurs with five input letters with the channel in Fig. 2. In this channel, it is possible to choose at most two non-adjacent letters, for example 0 and 2. Using sequences of these, 00, 02, 20, and 22 we obtain four words in a code of length two. However, it is possible to construct a code of length two with five members no two of which are adjacent as follows: 00, 12, 24, 31, 43. It is readily verified that no two of these are adjacent. Thus, C_0 for this channel is at least $\frac{1}{2} \log 5$.

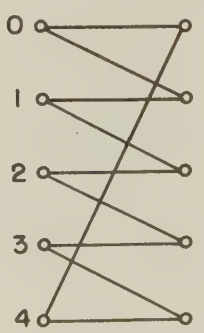


Fig. 2

No method has been found for determining C_0 for the general discrete channel, and this we propose as an interesting unsolved problem in coding theory. We shall develop a number of results which enable one to determine C_0 in many special cases, for example, in all channels with five or less input letters with the single exception of the channel of Fig. 2 (or channels equivalent in adjacency structure to it). We will also develop some general inequalities enabling one to estimate C_0 quite closely in most cases.

It may be seen, in the first place, that the value of C_0 depends only on which input letters are adjacent to each other. Let us define the adjacency matrix for a channel, A_{ij}, as follows.

$$A_{ij} = \begin{cases} 1 \text{ if input letter } i \text{ is adjacent to } j \text{ or} \\ \quad \text{if } i = j \\ 0 \text{ otherwise} \end{cases}$$

Suppose two channels have the same adjacency matrix (possibly after renumbering the input letters of one of them). Then it is obvious that a zero error code for one will be a zero error code for the other and, hence, that the zero error capacity C_0 for one will also apply to the other.

The adjacency structure contained in the adjacency matrix can also be represented as a linear graph. Construct a graph with as many vertices as there are input letters, and connect two distinct vertices with a line or branch of the graph if the corresponding input letters are adjacent. Two examples are shown in Fig. 3, corresponding to the channels of Figs. 1 and 2.

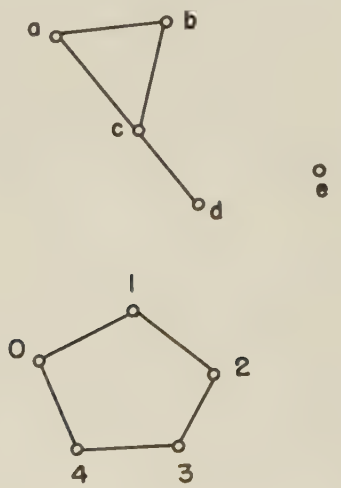


Fig. 3

Theorem 1: The zero error capacity C_0 of a discrete memoryless channel is bounded by the inequalities

$$-\log \min_{P_i} \sum_{ij} A_{ij} P_i P_j \leq C_0 \leq \min_{p_i(j)} C$$

$$\sum_i P_i = 1, \quad P_i \geq 0$$

$$\sum_j p_i(j) = 1, \quad p_i(j) \geq 0$$

where C is the capacity of any channel with transition probabilities $p_i(j)$ and having the adjacency matrix A_{ij}.

The upper bound is fairly obvious. The zero error capacity is certainly less than or equal to the ordinary capacity for any channel with the same adjacency matrix since the former requires codes with zero probability of error while the latter requires only codes approaching zero probability of error. By minimizing the capacity through variation of the $p_i(j)$ we find the lowest upper bound available through this argument. Since the capacity is a continuous function of the $p_i(j)$ in the closed region defined by $p_i(j) \geq 0$, $\sum_j p_i(j) = 1$, we may write min instead of greatest lower bound.

It is worth noting that it is only necessary to consider a particular channel in performing this minimization, although there are an infinite number with the same adjacency matrix. This one particular channel is obtained as follows from the adjacency matrix. If $A_{ik} = 1$ for a pair i k, define an output letter j with $p_i(j)$ and $p_k(j)$ both differing from zero. Now if there are any three input letters, say i k l, all adjacent to each other, define an output letter, say m, with $p_i(m)$ $p_k(m)$ $p_l(m)$ all different from zero. In the adjacency graph this corresponds to a complete sub-graph with three vertices. Next, subsets of four letters or complete subgraphs of four vertices, say i k l m, are given an output letter, each being connected to it, and so on. It is evident that any channel with the same adjacency matrix differs from that just described only by variation in the number of output symbols for some of the pairs, triplets, etc., of adjacent input letters. If a channel has more than one output symbol for an adjacent subset of input letters, then its capacity is reduced by identifying these. If a channel contains no element, say for a triplet i k l of adjacent input letters, this will occur as a special case of our canonical channel which has output letter m for this triplet when $p_i(m)$, $p_k(m)$ and $p_l(m)$ all vanish.

The lower bound of the theorem will now be proved. We use the procedure of random codes based on probabilities for the letters P_i, these being chosen to minimize the quadratic form $\sum_{ij} A_{ij} P_i P_j$. Construct an ensemble of codes

each containing M words, each word n letters long. The words in a code are chosen by the following stochastic method. Each letter of each word is chosen independently of all others and is the letter i with probability P_i. We now compute the probability in the ensemble that any particular word is not adjacent to any other word in its code. The probability that the first letter of one word is adjacent to the first letter of a second word is $\sum_{ij} A_{ij}P_iP_j$, since this sums the cases of adjacency with coefficient 1 and those of non-adjacency with coefficient 0. The probability that two words are adjacent in all letters, and therefore adjacent as words, is $(\sum_{ij} A_{ij}P_iP_j)^n$. The probability of non-adjacency is therefore $1 - (\sum_{ij} A_{ij}P_iP_j)^n$. The probability that all M - 1 other words in a code are not adjacent to a given word is, since they are chosen independently,

$$\left[1 - (\sum_{ij} A_{ij}P_iP_j)^n\right]^{M-1}$$

which is, by a well known inequality, greater than $1 - (M-1)(\sum_{ij} A_{ij}P_iP_j)^n$, which in turn is greater than $1 - M(\sum_{ij} A_{ij}P_iP_j)^n$. If we set $M = (1-\epsilon)^n(\sum_{ij} A_{ij}P_iP_j)^{-n}$, we then have, by taking ϵ small, a rate as close as desired to $-\log \sum_{ij} A_{ij}P_iP_j$. Furthermore, once ϵ is chosen, by taking n sufficiently large, we can insure that $M(\sum_{ij} A_{ij}P_iP_j)^n = (1-\epsilon)^n$ is as small as desired, say, less than δ. The probability in the ensemble of codes of a particular word being adjacent to any other in its own code is now less than δ. This implies that there are codes in the ensemble for which the ratio of the number of such undesired words to the total number in the code is less than or equal to δ. For, if not, the ensemble average would be worse than δ. Select such a code and delete from it the words having this property. We have reduced our rate only by at most $\log (1-\delta)^{-1}$. Since ϵ and δ were both arbitrarily small, we obtain error-free codes arbitrarily close to the rate-log $\min_{P_i} \sum_{ij} A_{ij}P_iP_j$ as stated in the theorem.

In connection with the upper bound of Theorem 1, the following result is useful in evaluating the minimum C. It is also interesting in its own right and will prove useful later in connection with channels having a feedback link.

Theorem 2: In a discrete memoryless channel with transition probabilities $p_i(j)$ and input letter probabilities P_i the following three statements are equivalent.

1) The rate of transmission

$$R = \sum_{i,j} P_ip_i(j) \log (p_i(j) / \sum_k P_kp_k(j))$$

is stationary under variation of all non-vanishing P_i subject to $\sum_i P_i = 1$ and under variation of $p_i(j)$ for those $p_i(j)$ such that $P_ip_i(j) > 0$ and subject to $\sum_j p_i(j) = 1$.

2) The mutual information between input-output pairs $I_{ij} = \log (p_i(j) / \sum_k P_kp_k(j))$ is constant, $I_{ij} = I$, for all ij pairs of non-vanishing probability (i.e. pairs for which $P_ip_i(j) > 0$).

3) We have $p_i(j) = r_j$ a function of j only whenever $P_ip_i(j) > 0$; and also $\sum_{i \in S_j} P_i = h$, a constant independent of j where S_j is the set of input letters that can produce output letter j with probability greater than zero. We also have $I = \log h^{-1}$.

The $p_i(j)$ and P_i corresponding to the maximum and minimum capacity when the $p_i(j)$ are varied (keeping, however, any $p_i(j)$ that are zero fixed at zero) satisfy 1), 2) and 3).

Proof: We will show first that 1) and 2) are equivalent and then that 2) and 3) are equivalent.

R is a bounded continouus function of its arguments P_i and $p_i(j)$ in the (bounded) region of allowed values defined by $\sum P_i = 1$, $P_i \geq 0$, $\sum_j p_i(j) = 1$, $p_i(j) \geq 0$. R has a finite partial derivative with respect to any $p_i(j) > 0$. In fact, we readily calculate

$$\frac{\partial R}{\partial p_i(j)} = P_i \log (p_i(j) / \sum_k P_kp_k(j))$$

A necessary and sufficient condition that R be stationary for small variation of the non-vanishing $p_i(j)$ subject to the conditions given is that

$$\frac{\partial R}{\partial p_i(j)} = \frac{\partial R}{\partial p_i(k)}$$

for all i, j, k such that P_i, $p_i(j)$, $p_i(k)$ do not vanish. This requires that

$$P_i \log p_i(j) / \sum_m P_mp_m(j) =$$

$$P_i \log p_i(k) / \sum_m P_mp_m(k)$$

If we let $Q_j = \sum_m P_mp_m(j)$, the probability of output letter j, then this is equivalent to

$$\frac{p_i(j)}{Q_j} = \frac{p_i(k)}{Q_k}$$

In other words, $p_i(j)/Q_j$ is independent of j, a function of i only whenever $P_i > 0$ and $p_i(j) > 0$. This function of i we call α_i. Thus

$$p_i(j) = \alpha_i Q_j$$

unless $P_i p_i(j) = 0$.

Now, taking the partial derivative of R with respect to P_i we obtain:

$$\frac{\partial R}{\partial P_i} = \sum_j p_i(j) \log \frac{p_i(j)}{Q_j} - 1$$

For R to be stationary subject to $\sum_i P_i = 1$ we must have $\partial R/\partial P_i = \partial R/\partial P_k$. Thus

$$\sum_j p_i(j) \log \frac{p_i(j)}{Q_j} = \sum_j p_k(j) \log \frac{p_k(j)}{Q_j}$$

Since for $P_i p_i(j) > 0$ we have $p_i(j)/Q_j = \alpha_i$, this becomes

$$\sum_j p_i(j) \log \alpha_i = \sum_j p_k(j) \log \alpha_k$$

$$\log \alpha_i = \log \alpha_k$$

Thus α_i is independent of i and may be written α. Consequently

$$\frac{p_i(j)}{Q_j} = \alpha$$

$$\log \frac{p_i(j)}{Q_j} = \log \alpha = I$$

whenever $P_i p_i(j) > 0$.

The converse result is an easy reversal of the above argument. If

$$\log \frac{p_i(j)}{Q_j} = I, \text{ then}$$

$\partial R/\partial P_i = I - 1$, by a simple substitution in the $\partial R/\partial P_i$ formula. Hence R is stationary under variation of P_i constrained by $\sum P_i = 1$. Further, $\partial R/\partial p_i(j) = P_i I = \partial R/\partial p_i(k)$, and hence the variation of R also vanishes subject to $\sum_j p_i(j) = 1$.

We now prove that 2) implies 3). Suppose $\log \frac{p_i(j)}{Q_j} = I$ whenever $P_i p_i(j) > 0$. Then $p_i(j) = e^I Q_j$, a function of j only under this same condition. Also, if $q_j(i)$ is the conditional probability of i given j, then

$$\frac{Q_j q_j(i)}{P_i Q_j} = e^I$$

$$q_j(i) = e^I P_i$$

$$1 = \sum_{i \in S_j} q_j(i) = e^I \sum_{i \in S_j} P_i$$

To prove that 3) implies 2) we assume

$$p_i(j) = r_j$$

when $P_i p_i(j) > 0$. Then

$$\frac{P_i p_i(j)}{P_i Q_j} = \frac{r_j}{Q_j} = \lambda_j \text{ (say)} = \frac{Q_j q_j(i)}{P_i Q_j} = \frac{q_j(i)}{P_i}$$

Now, summing the equation $P_i \lambda_j = q_j(i)$ over $i \in S_j$ and using the assumption from 3) that $\sum_{S_j} P_i = h$ we obtain

$$h \lambda_j = 1$$

so λ_j is h^{-1} and independent of j. Hence $I_{ij} = I = \log h^{-1}$.

The last statement of the theorem concerning minimum and maximum capacity under variation of $p_i(j)$ follows from the fact that R at these points must be stationary under variation of all non-vanishing P_i and $p_i(j)$, and hence the corresponding P_i and $p_i(j)$ satisfy condition 1) of the theorem.

For simple channels it is usually more convenient to apply particular tricks in trying to evaluate C_0 instead of the bounds given in Theorem 1, which involve maximizing and minimizing processes. The simplest lower bound, as mentioned before, is obtained by merely finding the logarithm of the maximum number of non-adjacent input letters.

A very useful device for determining C_0 which works in many cases may be described using the notion of an adjacency-reducing mapping.

By this we mean a mapping of letters into other letters, $i \to \alpha(i)$, with the property that if i and j are not adjacent in the channel (or graph) then $\alpha(i)$ and $\alpha(j)$ are not adjacent. If we have a zero-error code, then we may apply such a mapping letter by letter to the code and obtain a new code which will also be of the zero-error type, since no adjacencies can be produced by the mapping.

Theorem 3: If all the input letters i can be mapped by an adjacency-reducing mapping $i \to \alpha(i)$ into a subset of the letters no two of which are adjacent, then the zero-error capacity C_0 of the channel is equal to the logarithm of the number of letters in this subset.

For, in the first place, by forming all sequences of these letters we obtain a zero-error code at this rate. Secondly, any zero error code for the channel can be mapped into a code using only these letters and containing, therefore, at most $e^{C_0 n}$ non-adjacent words.

The zero-error capacities, or, more exactly, the equivalent numbers of input letters for all adjacency graphs up to five vertices are shown in Fig. 4. These can all be found readily by the method of Theorem 3, except for the channel of Fig. 2 mentioned previously, for which we know only that the zero-error capacity lies in the range $\frac{1}{2} \log 5 \leq C_0 \leq \log \frac{5}{2}$.

All graphs with six vertices have been examined and the capacities of all of these can also be found by this theorem, with the exception of four. These four can be given in terms of the capacity of Fig. 2, so that this case is essentially the only unsolved problem up to seven vertices. Graphs with seven vertices have not been completely examined but at least one new situation arises, the analog of Fig. 2 with seven input letters.

As examples of how the N_0 values were computed by the method of adjacency-reducing mappings, several of the graphs in Fig. 4 have been labelled to show a suitable mapping. The scheme is as follows. All nodes labelled a are mapped into node α as well as α itself. All nodes labelled b and also β are mapped into node β. All nodes labelled c and γ are mapped into node γ. It is readily verified that no new adjacencies are produced by the mappings indicated and that the α, β, γ nodes are non-adjacent.

C_0 for Sum and Product Channels

Theorem 4: If two memoryless channels have zero-error capacities $C_0' = \log A$ and $C_0'' = \log B$, their sum has a zero-error capacity greater than or equal to $\log (A + B)$ and their product a zero error capacity greater than or equal to $C_0' + C_0''$. If the graph of either of the two channels can be reduced to non-adjacent points by the mapping method (Theorem 3), then these inequalities can be replaced by equalities.

Proof: It is clear that in the case of the product, the zero error capacity is at least $C_0' + C_0''$, since we may form a product code from two codes with rates close to C_0' and C_0''. If these codes are not of the same length, we use for the new code the least common multiple of the individual lengths and form all sequences of the code words of each of the codes up to this length. To prove equality in case one of the graphs, say that for the first channel, can be mapped into A non-adjacent points, suppose we have a code for the product channel. The letters for the product code, of course, are ordered pairs of letters corresponding to the original channels. Replace the first letter in each pair in all code words by the letter corresponding to reduction by the mapping method. This reduces or preserves adjacency between words in the code. Now sort the code words into A^n subsets according to the sequences of first letters in the ordered pairs. Each of these subsets can contain at most B^n members, since this is the largest possible number of codes for the second channel of this length. Thus, in total, there are at most $A^n B^n$ words in the code, giving the desired result.

In the case of the sum of the two channels, we first show how, from two given codes for the two channels, to construct a code for the sum channel with equivalent number of letters equal to $A^{1-\delta} + B^{1-\delta}$, where δ is arbitrarily small and A and B are the equivalent number of letters for the two codes. Let the two codes have lengths n_1 and n_2. The new code will have length n where n is the smallest integer greater than both $\frac{n_1}{\delta}$ and $\frac{n_2}{\delta}$. Now form codes for the first channel and for the second channel for all lengths k from zero to n as follows. Let k equal $an_1 + b$, where a and b are integers and $b < n_1$. We form all sequences of a words from the given code for the first channel and fill in the remaining b letters arbitrarily, say all with the first letter in the code alphabet. We achieve at least $A^{k - \delta n}$ different words of length k none of which is adjacent to any other. In the same way we form codes for the second channel and achieve $B^{k - \delta n}$ words in this code of length k. We now intermingle the k code for the first channel with the n - k code for the second channel in all $\binom{n}{k}$ possible ways and do this for each value of k. This produces a code n letters long with at least

$$\sum_{k=0}^{n} \binom{n}{k} A^{k - n\delta} B^{n - k - n\delta}$$

$$= (AB)^{-\delta n}(A + B)^n$$ different words. It is readily seen that no two of these different words are adjacent. The rate is at least $\log (A + B) - \delta \log AB$, and since δ was arbitrarily small, we can achieve a rate arbitrarily close to $\log (A + B)$.

To show that it is not possible, when one of the graphs reduces by mapping to non-adjacent points, to exceed the rate corresponding to the number of letters A + B, consider any given code of length n for the sum channel. The words in this consist of sequences of letters each letter corresponding to one or the other of the two

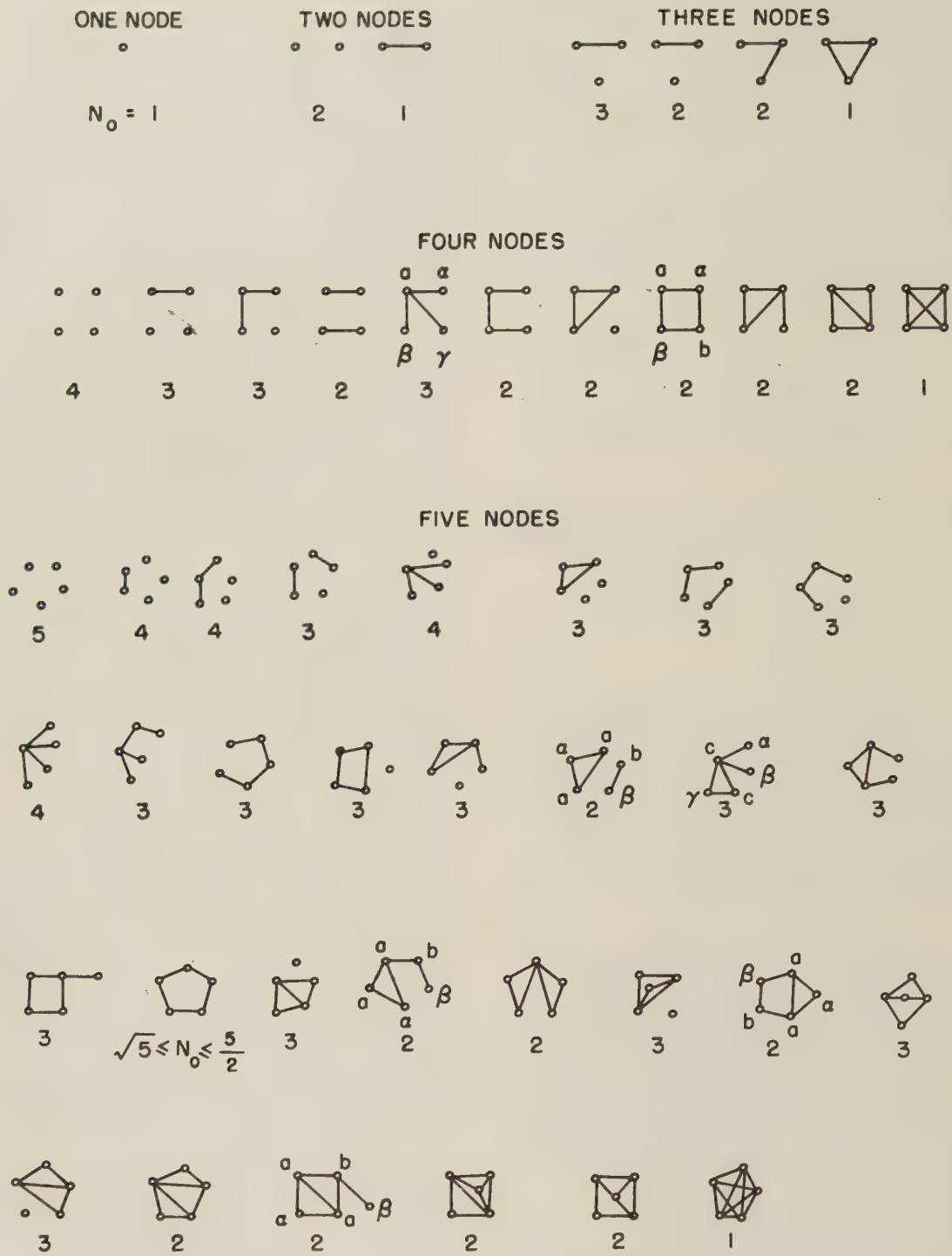


Fig. 4 - All graphs with 1, 2, 3, 4, 5 nodes and the corresponding N_0 for channels with these as adjacency graphs (note $C_0 = \log N_0$)

channels. The words may be subdivided into classes corresponding to the pattern of the choices of letters between the two channels. There are 2^n such classes with $\binom{n}{k}$ classes in which exactly k of the letters are from the first channel and n - k from the second. Consider now a particular class of words of this type. Replace the letters from the first channel alphabet by the corresponding non-adjacent letters. This does not harm the adjacency relations between words in the code. Now, as in the product case, partition the code words according to the sequence of letters involved from the first channel. This produces at most A^k subsets. Each of these subsets contains at most B^{n-k} members, since this is the greatest possible number of non-adjacent words for the second channel of length n - k. In total, then, summing over all values of k and taking account of the $\binom{n}{k}$ classes for each k, there are at most $\sum_k \binom{n}{k} A^k B^{n-k}$

$=(A + B)^n$ words in the code for the sum channel. This proves the desired result.

Theorem 4, of course, is analogous to known results for ordinary capacity C, where the product channel has the sum of the ordinary capacities and the sum channel has an equivalent number of letters equal to the sum of the equivalent numbers of letters for the individual channels. We conjecture but have not been able to prove that the equalities in Theorem 4 hold in general, not just under the conditions given. We now prove a lower bound for the probability of error when transmitting at a rate greater than C_0.

Theorem 5: In any code of length n and rate $R > C_0$, $C_0 > 0$, the probability of error P_e will satisfy $P_e \geq (1 - e^{-n(C_0 - R)}) p_{min}^n$, where p_{min} is the minimum non-vanishing $p_i(j)$.

Proof: By definition of C_0 there are not more than e^{nC_0} non-adjacent words of length n. With $R > C_0$, among e^{nR} words there must, therefore, be an adjacent pair. The adjacent pair has a common output word which either can cause with a probability at least p_{min}^n. This output word cannot be decoded into both inputs. At least one, therefore, must cause an error when it leads to this output word. This gives a contribution at least $e^{-nR} p_{min}^n$ to the probability of error P_e. Now omit this word from consideration and apply the same argument to the remaining $e^{nR} - 1$ words of the code. This will give another adjacent pair and another contribution of error of at least $e^{-nR} p_{min}^n$. The process may be continued until the number of code points remaining is just e^{nC_0}. At this time, the computed probability of error must be at least $(e^{nR} - e^{nC_0}) e^{-nR} p_{min}^n$

$= (1 - e^{n(C_0 - R)}) p_{min}^n$.

Channels with a Feedback Link

We now consider the corresponding problem for channels with complete feedback. By this we mean that there exists a return channel sending back from the receiving point to the transmitting point, without error, the letters actually received. It is assumed that this information is received at the transmitting point before the next letter is transmitted, and can be used, therefore, if desired, in choosing the next transmitted letter.

It is interesting that for a memoryless channel the ordinary forward capacity is the same with or without feedback. This will be shown in Theorem 6. On the other hand, the zero error capacity may, in some cases, be greater with feedback than without. In the channel shown in Fig. 5, for example, $C_0 = \log 2$. However, we will see as a result of Theorem 7 that with feedback the zero error capacity $C_{0F} = \log 2.5$.

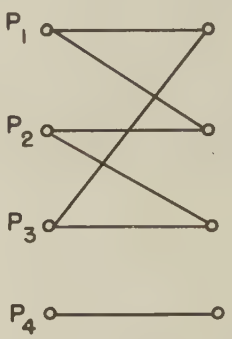


Fig. 5

We first define a block code of length n for a feedback system. This means that at the transmitting point there is a device with two inputs, or, mathematically, a function with two arguments. One argument is the message to be transmitted, the other, the past received letters (which have come in over the feedback link). The value of the function is the next letter to be transmitted. Thus, the function may be thought of as $x_{j+1} = f(k, v_j)$ where x_{j+1} is the $j+1$ transmitted letter in a block, k is an index ranging from 1 to M, and represents the specific message, and v_j is a received word of length j. Thus j ranges from 0 to n - 1 and v_j over all received words of these lengths.

In operation, if message m_k is to be sent f is evaluated for f(k -) where the - means "no

word" and this is sent as the first transmitted letter. If the feedback link sends back α, say, as the first received letter, the next transmitted letter will be $f(k, \alpha)$. If this is received as β. the next transmitted letter will be $f(k, \alpha\beta)$, etc.

Theorem 6: In a memoryless discrete channel with feedback, the forward capacity is equal to the ordinary capacity C (without feedback). The average change in mutual information I_{vm} between received sequence v and message m for a letter of text is not greater than C.

Proof: Let v be the received sequence to date of a block, m the message, x the next transmitted letter and y the next received letter. These are all random variables and, also, x is a function of m and v. This function, namely, is the one which defines the encoding procedure with feedback whereby the next transmitted letter x is determined by the message m and the feedback information v from the previous received signals. The channel being memoryless implies that the next operation is independent of the past, in particular, $Pr[y/x] = Pr[y/x,v]$.

The average change in mutual information, when a particular v has been received, due to the x,y pair is given by (we are averaging over messages m and next received letters y, for a given v):

$$\overline{\Delta I} = \overline{I_{m,vy}} - \overline{I_{m,v}} = \sum_{y,m} Pr[y,m/v] \cdot \log \frac{Pr[v,y,m]}{Pr[v,y]Pr[m]} - \sum_{m} Pr[m/v] \cdot \log \frac{Pr[v,m]}{Pr[v]Pr[m]}$$

Since $Pr[m/v] = \sum_{y} Pr[y,m/v]$, the second sum may be rewritten as $\sum_{y,m} Pr[y,m/v] \log \frac{Pr[v,m]}{Pr[v]\,Pr[m]}$

The two sums then combine to give

$$\overline{\Delta I} = \sum_{y,m} Pr[y,m/v] \log \frac{Pr[v,y,m]\,Pr[v]}{Pr[v,m]\,Pr[v,y]}$$

$$= \sum_{y,m} Pr[y,m/v] \log \frac{Pr[y/v,m]\,Pr[v]}{Pr[v,y]}$$

The sum on m may be thought of as summed first on the m's which result in the same x (for the given v), recalling that x is a function of m and v, and then summing on the different x's. In the first summation, the term $Pr[y/v,m]$ is constant at $Pr[y/x]$ and the coefficient of the logarithm sums to $Pr[x,y/v]$. Thus we can write

$$\Delta I = \sum_{x,y} Pr[x,y/v] \log \frac{Pr[y/x]}{Pr[y/v]}$$

Now consider the rate for the channel (in the ordinary sense without feedback) if we should assign to the x's the probabilities $q(x) = Pr[x/v]$. The probabilities for pairs, $r(x,y)$, and for the y's alone, $w(y)$, in this situation would then be

$$r(x,y) = q(x)\, Pr[y/x] = Pr[x/v]\, Pr[y/x] = Pr[x,y/v]$$

$$w(y) = \sum_{x} r(x,y) = \sum_{x} Pr[x,y/v] = Pr[y/v]$$

Hence the rate would be

$$R = \sum_{x,y} r(x,y) \log \frac{Pr[y/x]}{w(y)} = \sum_{x,y} Pr[x,y/v] \log \frac{Pr[y/x]}{Pr[y/v]} = \Delta I$$

Since $R \leq C$, the channel capacity (C being the maximum possible R for all q(x) assignments), we conclude that

$$\Delta I \leq C.$$

Since the average change in I per letter is not greater than C, the average change in n letters is not greater than nC. Hence, in a block code of length n with input rate R, if $R > C$ then the equivocation at the end of a block will be at least R - C, just as in the non-feedback case. In other words, it is not possible to approach zero equivocation (or, as easily follows, zero probability of error) at a rate exceeding the channel capacity. It is, of course, possible to do this at rates less than C, since certainly anything that can be done without feedback can be done with feedback.

It is interesting that the first sentence of Theorem 6 can be generalized readily to channels with memory provided they are of such a nature that the internal state of the channel can be calculated at the transmitting point from the initial state and the sequence of letters that have been transmitted. If this is not the case, the conclusion of the theorem will not always be true, that is, there exist channels of a more complex sort for which the forward capacity with feedback exceeds that without feedback. We shall not, however, give the details of these generalizations here.

Returning now to the zero-error problem, we define a zero error capacity C_{oF} for a channel with feedback in the obvious way--the least upper bound of rates for block codes with no errors. The next theorem solves the problem of evaluating C_{oF} for memoryless channels with feedback, and indicates how rapidly C_{oF} may be approached as the block length n increases.

Theorem 7: In a memoryless discrete channel with complete feedback of received letters to the transmitting point, the zero error capacity C_{oF} is zero if all pairs of input letters are adjacent. Otherwise $C_{oF} = \log P_o^{-1}$ where

$$P_o = \min_{P_i} \max_{j} \sum_{i \in S_j} P_i$$

P_i being a probability assigned to input letter i ($\sum_i P_i = 1$) and S_j the set of input letters which can cause output letter j with probability greater than zero. A zero error block code of length n can be found for such a feedback channel which transmits at a rate $R \geq C_{oF}\left(1 - \frac{2}{n} \log_2 2t\right)$ where t is the number of input letters.

The P_o occuring in this theorem has the following meaning. For any given assignment of probabilities P_i to the input letters one may calculate, for each output letter j, the total probability of all input letters that can (with positive probability) cause j. This is $\sum_{i \in S_j} P_i$. Output letters for which this is large may be thought of as "bad" in that when received there is a large uncertainty as to the cause. To obtain P_o one adjusts the P_i so that worst output letter in this sense is as good as possible.

We first show that if all letters are adjacent to each other $C_{oF} = 0$. In fact, in any coding system, any two messages, say m_1 and m_2 can lead to the same received sequence with positive probability. Namely, the first transmitted letters corresponding to m_1 and m_2 have a possible received letter in common. Assuming this occurs, calculate the next transmitted letters in the coding system for m_1 and m_2. These also have a possible received letter in common. Continuing in this manner we establish a received word which could be produced by either m_1 or m_2 and therefore they cannot be distinguished with certainty.

Now consider the case where not all pairs are adjacent. We will first prove, by induction on the block length n, that the rate $\log P_o^{-1}$ cannot be exceeded with a zero error code. For n = o the result is certainly true. The inductive hypothesis will be that no block code of length n - 1 transmits at a rate greater than $\log P_o^{-1}$, or, in other words, can resolve with certainty more than

$$e^{(n-1) \log P_o^{-1}} = P_o^{-(n-1)}$$

different messages. Now suppose (in contradiction to the desired result) we have a block code of length n resolving M messages with $M > P_o^{-n}$. The first transmitted letter for the code partitions these M messages among the input letters for the channel. Let F_i be the fraction of the messages assigned to letter i (that is, for which i is the first transmitted letter). Now these F_i are like probability assignments to the different letters and therefore by definition of P_o, there is some output letter, say letter k, such that $\sum_{i \in S_k} F_i \geq P_o$. Consider the set of messages for which the first transmitted letter belongs to S_k. The number of messages in this set is at least $P_o M$. Any of these can cause output letter k as first received letter. When this happens there are n - 1 letters yet to be transmitted and since $M > P_o^{-n}$ we have $P_o M > P_o^{-(n-1)}$. Thus we have a zero error code of block length n - 1 transmitting at a rate greater than $\log P_i^{-1}$, contradicting the inductive assumption. Note that the coding function for this code of length n - 1 is formally defined from the original coding function by fixing the first received letter at k.

We must now show that the rate $\log P_o^{-1}$ can actually be approached as closely as desired with zero error codes. Let P_i be the set of probabilities which, when assigned to the input letters, give P_o for $\min_{P_i} \max_{j} \sum_{i \in S_j} P_i$. The general scheme of the code will be to divide the M original messages into t different groups corresponding to the first transmitted letter. The number of messages in these groups will be approximately proportional to $P_1, P_2, \ldots P_t$. The first transmitted letter, then, will correspond to the group containing the message to be transmitted. Whatever letter is received, the number of possible messages compatible with this

received letter will be approximately $P_o M$. This subset of possible messages is known both at the receiver and (after the received letter is sent back to the transmitter) at the transmitting point.

The code system next subdivides this subset of messages into t groups, again approximately in proportion to the probabilities P_i. The second letter transmitted is that corresponding to the group containing the actual message. Whatever letter is received, the number of messages compatible with the two received letters is now, roughly, $P_o^2 M$.

This process is continued until only a few messages (less than t^2) are compatible with all the received letters. The ambiguity among these is then resolved by using a pair of non-adjacent letters in a simple binary code. The code thus constructed will be a zero error code for the channel.

Our first concern is to estimate carefully the approximation involved in subdividing the messages into the t groups. We will show that for any M and any set of P_i $\sum P_i = 1$, it is possible to subdivide the M messages into groups of $m_1, m_2, \ldots m_t$ such that $m_i = o$ whenever $P_i = o$ and

$$\left| \frac{m_i}{M} - P_i \right| \leq \frac{1}{M} \qquad i = 1, \ldots, t$$

We assume without loss of generality that $P_1, P_2, \ldots P_s$ are the non-vanishing P_i. Choose m_1 to be the largest integer such that $\frac{m_1}{M} \leq P_1$. Let $P_1 - \frac{m_1}{M} = \delta_1$. Clearly $|\delta_1| \leq \frac{1}{M}$. Next choose m_2 to be the smallest integer such that $\frac{m_2}{M} \geq P_2$ and let $P_2 - \frac{m_2}{M} = \delta_2$. We have $|\delta_2| \leq \frac{1}{M}$. Also $|\delta_1 + \delta_2| \leq \frac{1}{M}$, since δ_1 and δ_2 are opposite in sign and each less than $\frac{1}{M}$ in absolute value. Next, m_3 is chosen so that $\frac{m_3}{M}$ approximates, to within $\frac{1}{M}$, to P_3. If $\delta_1 + \delta_2 \geq 0$, then $\frac{m_3}{M}$ is chosen less than or equal to P_3. If $\delta_1 = \delta_2 < 0$, then $\frac{m_3}{M}$ is chosen greater than or equal to P_3. Thus again $P_3 - \frac{m_3}{M} = \delta_3 \leq \frac{1}{M}$ and $|\delta_1 + \delta_2 + \delta_3| \leq \frac{1}{M}$. Continuing in this manner through P_{s-1} we obtain approximations for $P_1, P_2, \ldots, P_{s-1}$ with the property that

$$\left| \delta_1 + \delta_2 + \ldots + \delta_{s-1} \right| \leq \frac{1}{M}, \text{ or}$$

$$\left| M (P_1 + P_2 + \ldots + P_{s-1}) - (m_1 + m_2 + \ldots + m_{s-1}) \right| \leq 1.$$

If we now define m_s as $M - \sum_1^{s-1} m_i$ then this inequality can be written $\left| M(1 - P_s) - (M - m_s) \right| \leq 1$. Hence $\left| \frac{m_s}{M} - P_s \right| \leq \frac{1}{M}$. Thus we have achieved the objective of keeping all approximation $\frac{m_i}{M}$ to within $\frac{1}{M}$ of P_i and having $\sum m_i = M$.

Returning now to our main problem note first that if $P_o = 1$ then $C_{oF} = 0$ and the theorem is trivially true. We assume, then, that $P_o < 1$. We wish to show that $P_o \leq (1 - \frac{1}{t})$. Consider the set of input letters which have the maximum value of P_i. This maximum is certainly greater than or equal to the average $\frac{1}{t}$. Furthermore, we can arrange to have at least one of these input letters not connected to some output letter. For suppose this is not the case. Then either there are no other input letters beside this set and we contradict the assumption that $P_o < 1$, or there are other input letters with smaller values of P_i. In this case, by reducing the P_i for one input letter in the maximum set and increasing correspondingly that for some input letter which does not connect to all output letters, we do not increase the value of P_o (for any S_j) and create an input letter of the desired type. By consideration of an output letter to which this input letter does not connect we see that $P_o \leq 1 - \frac{1}{t}$.

Now suppose we start with M messages and subdivide into groups approximating proportionality to the P_i as described above. Then when a letter has been received, the set of possible messages (compatible with this received letter) will be reduced to those in the groups corresponding to letters which connect to the actual received letter. Each output letter connects to not more than t - 1 input letters (otherwise we would have $P_o = 1$). For each of the connecting groups, the error in approximating P_i has been less than or equal to $\frac{1}{M}$. Hence the total relative number in all connecting groups for any output letter is less than or equal to $P_o + \frac{t-1}{M}$. The total number of possible messages after receiving the first letter consequently drops from M to a number less than or equal to $P_o M + t - 1$.

In the coding system to be used, this remaining possible subset of messages is subdivided again among the input letters to approximate in the same fashion the probabilities P_i. This subdivision can be carried out both at

receiving point and transmitting point using the same standard procedure (say, exactly the one described above) since with the feedback both terminals have available the required data, namely the first received letter.

The second transmitted letter obtained by this procedure will again reduce at the receiving point the number of possible messages to a value not greater than $P_0 (P_0 M + t - 1) + t - 1$. This same process continues with each transmitted letter. If the upper bound on the number of possible remaining messages after k letters is M_k, then $M_{k+1} = P_0 M_k + t - 1$. The solution of this difference equation is

$$M_k = AP_0^k + \frac{t-1}{1-P_0}$$

This may be readily verified by substitution in the difference equation. To satisfy the initial conditions $M_0 = M$ requires $A = M - \frac{t-1}{1-P_0}$. Thus the solution becomes

$$M_k = (M - \frac{t-1}{1-P_0}) P_0^k + \frac{t-1}{1-P_0}$$

$$= M P_0^k + \frac{t-1}{1-P_0}(1 - P_0^k)$$

$$\leq M P_0^k + t(t-1)$$

since we have seen above that $1 - P_0 \geq \frac{1}{t}$.

If the process described is carried out for n_1 steps, where n_1 is the smallest integer $\geq d$ where d is the solution of $MP_0^d = 1$, then the number of possible messages left consistent with the received sequence will be not greater than $1 + t(t-1) \leq t^2$ (since $t > 1$, otherwise we should have $C_{0F} = 0$). Now the pair of non-adjacent letters assumed in the theorem may be used to resolve the ambiguity among these t^2 or less messages. This will require not more than $1 + \log_2 t^2 = \log_2 2t^2$ additional letters. Thus, in total, we have used not more than $d + 1 + \log_2 2t^2 = d + \log_2 4t^2 = n$ say as block length. We have transmitted in this block length a choice from $M = P_0^{-d}$ messages. Thus the zero error rate we have achieved is

$$R = \frac{1}{n} \log M \geq \frac{d \log P_0^{-1}}{d + \log_2 4t^2}$$

$$= (1 - \frac{1}{n} \log 4t^2) \log P_0^{-1}$$

$$= (1 - \frac{1}{n} \log 4t^2) \; C_{0F}$$

Thus we can approximate to C_{0F} as closely as desired with zero error codes.

As an example of Theorem 7 consider the channel in Fig. 5. We wish to evaluate P_0. It is easily seen that we may take $P_1 = P_2 = P_3$ in forming the min max of Theorem 7, for if they are unequal the maximum $\sum_{i \epsilon S_j} P_i$ for the corresponding three output letters would be reduced by equalizing. Also it is evident, then, that $P_4 = P_1 + P_2$, since otherwise a shift of probability one way or the other would reduce the maximum. We conclude, then, that $P_1 = P_2 = P_3 = 1/5$ and $P_4 = 2/5$. Finally, the zero error capacity with feedback is $\log P_0^{-1} = \log 5/2$.

There is a close connection between the min max process of Theorem 7 and the process of finding the minimum capacity for the channel under variation of the non-vanishing transition probabilities $p_i(j)$ as in Theorem 2. It was noted there that at the minimum capacity each output letter can be caused by the same total probability of input letters. Indeed, it seems very likely that the probabilities of input letters to attain the minimum capacity are exactly those which solve the min max problem of Theorem 7, and, if this is so, the $C_{min} = \log P_0^{-1}$.

Acknowledgement

I am indebted to Peter Elias for first pointing out that a feedback link could increase the zero-error capacity, as well as for several suggestions that were helpful in the proof of Theorem 7.

A LINEAR CIRCUIT VIEWPOINT ON ERROR-CORRECTING CODES*

David A. Huffman
Department of Electrical Engineering
and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract

A linear binary filter has as its output a binary sequence, each digit of which is the result of a parity check on a selection of preceding output digits and of present and preceding digits of the filter input sequence. The terminal properties of these filters may be described by transfer ratios of polynomials in a delay operator. If two binary filters have transfer ratios which are reciprocally related then the filters are mutually inverse in the sense that, in a cascade connection, the second filter unscrambles the scrambling produced by the first.

The coding of a finite sequence of binary information digits for protection against noise may be accomplished by a binary sequence filter, the output of which becomes the sequence to be transmitted. (The inverse filter is utilized at the receiver.) Into the filter at the transmitter is inserted a sequence of information digits, immediately followed by another sequence of completely predictable digits consisting, say, of zeros. The completed block of digits is scrambled in a linear filter before transmission through the noisy channel. If this scrambled sequence were unaffected by noise in the channel the result of unscrambling by the receiver filter would be the original sequence of information digits followed by the all-zero sequence. If, however, channel noise has been added to the sequence put into the receiver filter, then its output is the original sequence, plus the response of the receiver filter to the noise superimposed thereon. In particular, the sequence positions which would have contained all zeros, had there been no noise, will now contain digits whose values are related to the sequence positions affected by the noise. These data may then be utilized for the subsequent correction of the errors which would otherwise have been caused by the noise.

I Algebraic Description and Realization of Linear Sequence Filters

A linear binary sequence filter[1] is a synchronous filter whose inputs and outputs are ordered sequences of binary symbols (0's and 1's). For the general non-time-varying filter each digit of the filter output sequence is a modulo-two sum of an arbitrary selection of past output digits (Z) and of present and past input digits (X). The description of a sequence filter in terms of a delay operator, D, is a straightforward one. For example, a filter whose output Z is the sum of the first and third previous output digits and of the present, first, second, and fourth previous input digits is described by

$$Z = DZ + D^3Z + X + DX + D^2X + D^4X \qquad (1)$$

where the + symbol is used here for the modulo-two operation. That is, the present output is zero if an even number of selected digits have the value one, and is unity if an odd number have the value one.

Since the modulo-two operation is self-inverse the terms in (1) may be rearranged to give

$$D^3Z + DZ + Z = D^4X + D^2X + DX + X \qquad (2\text{-a})$$

or

$$(D^3 + D + I)Z = (D^4 + D^2 + D + I)X \qquad (2\text{-b})$$

The "transfer ratio" of the filter is then

$$\frac{Z}{X} = \frac{D^4 + D^2 + D + I}{D^3 + D + I} \qquad (3)$$

An efficient realization of this filter results from rearranging (1) to give

$$X + Z = D(X+Z) + D^2X + D^3Z + D^4X \qquad (4\text{-a})$$

or

$$X + Z = D\left\{(X+Z) + D\left\{X + D\left\{Z + DX\right\}\right\}\right\} \qquad (4\text{-b})$$

*This work was supported in part by the Signal Corps, the Office of Scientific Research (Air Research and Development Command), and the Office of Naval Research, of the United States.

The corresponding filter is given in Fig. 1-a. The "inverse" filter, whose input is Z and whose output is X is described by Eqs. 4-a,b and has a transfer ratio

$$\frac{X}{Z} = \frac{D^3 + D + I}{D^4 + D^2 + D + I} \qquad (5)$$

Its realization is given in Fig. 1-b. Both of the filters in Fig. 1 utilize only two kinds of elements: modulo-two adders and unit delays (single-stage shift-registers). The "chain" realization given both of these filters consists of a chain of unit delays with provision made for introducing the signals X, Z or (X+Z) between each two stages of delay. It uses just the number of delay units necessary to remember the input or output digit most remote in the past which is needed for proper operation of the filter (in this case the fourth previous input), and an equal number of adders.

When a binary filter and its inverse are connected in cascade one mode of operation of the combination is that for which the transfer ratio is the identity operator. In our example

$$\left(\frac{D^4 + D^2 + D + I}{D^3 + D + I}\right) \cdot \left(\frac{D^3 + D + I}{D^4 + D^2 + D + I}\right) = I \qquad (6)$$

That is, the second filter unscrambles the scrambling produced by the first. In the error-correcting scheme proposed in this paper the use of filters and their inverses will be of paramount importance.

II Description of a Filter From Its Impulse Response Characteristics

Later in this paper we will want to make use of the fact that there exist finite realizations of linear sequence filters whose response to an input "impulse" (a single digit 1 preceded and followed by infinite sequences of 0's) is arbitrary except that it must eventually die out (become the all-0 sequence) or ultimately become periodic. Suppose, for example, that we wish to realize a filter whose response to an input sequence, X*, containing an impulse is the output sequence, Z*, which ultimately becomes periodic. (See Fig. 2-a.) Z* can always be considered to be the sum of two sequences: Z^*_p, the periodic component, and Z^*_t the transient component. The filter we are trying to design may, for the moment, be considered to be made up of two sub-filters, f_p and f_t, which have impulse responses Z^*_p and Z^*_t, respectively, and whose outputs are added to give the desired response Z* (see Fig. 2-b).

The filter f_p could be realized by a cascade of two other filters (see Fig. 2-c). The first would have an impulse response which consisted of a sequence of impulses spaced seven intervals apart (the period of the periodic response) and continuing indefinitely. This filter would have a transfer ratio

$$I + D^7 + D^{14} + D^{21} + \cdots = \frac{I}{D^7 + I} \qquad (7)$$

The periodically recurring output of this filter could be used as the input to another filter having the proper transient response (finite in length). The latter filter has a transfer ratio which is a polynomial, $I + D + D^2 + D^4$ in this example, whose terms correspond to the positions of the 1's in a typical cycle, contained between commas, of the desired Z^*_p.

The transient part, Z^*_t, of the impulse response is easy to arrange for in our example. The proper associated filter, f_t, has a transfer ratio D.

The filter we are designing could then be realized with a total transfer ratio of

$$\frac{Z^*}{X^*} = \left(\frac{I}{D^7 + I}\right)(I + D + D^2 + D^4) + D \qquad (8\text{-a})$$

which may be rewritten as

$$\frac{Z^*}{X^*} = \frac{(I + D + D^2 + D^4) + D(D^7 + I)}{D^7 + I} \qquad (8\text{-b})$$

or as

$$\frac{Z^*}{X^*} = \frac{D^8 + D^4 + D^2 + I}{D^7 + I} \qquad (8\text{-c})$$

The numerator and denominator of the expression in Eq. 8-c each contain the factor $D^4 + D^2 + D + I$ (found using the Euclidean Algorithm; see reference (1)) which may be cancelled to give

$$\frac{Z^*}{X^*} = \frac{(D^4+D^2+D+I)(D^4+D^2+D+I)}{(D^4+D^2+D+I)(D^3+D+I)} = \frac{D^4+D^2+D+I}{D^3+D+1} \qquad (8\text{-d})$$

The transfer ratio is now in its simplest form and the filter may be synthesized as has already been done in Eqs. 4 and Fig. 1.

III A Linear Single-Error Correcting Coding Scheme

Consider the arrangement of filters shown in Fig. 3-a. A sequence of seven X digits, is fed into a transmitter filter with transfer ratio, T, resulting in a sequence Z = (T)X which is transmitted through the noisy channel. In the channel a noise sequence, N, is added to Z so that what arrives at the receiver filter is

$$Z' = Z + N \qquad (9\text{-a})$$

At the receiver a filter inverse to the transmitter filter creates from the sequence Z' a sequence

$$\begin{aligned} X' &= (T^{-1})Z' = (T^{-1})\left[Z + N\right] \\ &= (T^{-1})\left[(T)X + N\right] = X + (T^{-1})N \end{aligned} \qquad (9\text{-b})$$

If there were no noise in the channel (N = 0), X = X'. If there is noise present then the

sequence X' contains the sum of the transmitter input sequence, X, and the response $(T^{-1})N$ of the receiver filter to the noise. If only a single noise digit is present the sequence X' contains X plus the impulse response of the receiver filter superimposed thereon.

Let us examine the coding and decoding mechanism in more detail. The first four digits of the sequence X are information digits, and may therefore be chosen in $2^4 = 16$ different ways. The remaining three digits are always all zeros and are to be called here buffer digits. The composite block of seven digits is scrambled for transmission in the channel by the first filter. The sequence X' which results from the unscrambling action of the receiver filter would equal X if there were no noise in the channel. The clue to this possibility would be the existence of three zeros in the buffer positions (the last three digits) in the sequence X'.

When a single noise digit is equal to unity (just one transmitted digit is changed by noise action) the received sequence, X', may look quite different from X. (See Fig. 3-b.) In particular the buffer positions will no longer contain all zeros, but will instead be three successive digits of the impulse response of the receiver filter. In our example the impulse response of that filter is given in Fig. 3-d, and since we have assumed the noise impulse to occur in the third position of the block of seven digits we observe in the buffer positions the third, fourth, and fifth digits of the impulse response. (See Fig. 3-c,d).

It is extremely important to notice that the digits in the buffer positions of the sequence X' are independent of which of the sixteen possible X sequences is sent. This pattern of digits depends only upon the position(s) of the noise digits and upon the impulse response of the receiver filter.

We have chosen the receiver filter so that its impulse response has a period of seven digits (the length of the composite block) and so that each of the seven possible combinations of three (the number of buffer digits and the degree of the denominator polynomial) successive digits in the response will be different from the others. That this is possible for a block length of $n = 2^b - 1$ with b buffer positions follows from the fact that the maximum possible period of the impulse response of a filter with denominator polynomial of degree b is $2^b - 1$. (See reference 1.)

By observing the three buffer positions of the sequence X', and by knowing the form of the impulse response of the receiver filter we can deduce where the noise impulse occurred in the block. If we assume that only a single noise impulse was present (the most likely situation) we can recreate the original sequence X by adding (same as subtracting, modulo-two) the now known sequence $(T^{-1})N$ to the sequence X'.

For our example it is interesting to examine the sixteen possible sequences, Z, which correspond to the sixteen possible sequences, X, which might be inserted into the transmitter filter with $T = D^3 + D^2 + I$. These are listed in Fig. 4. The sixteen Z sequences are mutually separated by a distance of at least three, a necessary condition for single-error correction[2].

The advantage of the linear circuit viewpoint of this paper is that instead of concerning ourselves with the distance properties of $2^k = 2^{n-b}$ (in our example, 16) different code message sequences $Z = (T)X$, we may concentrate our attention on the impulse response of the receiver filter with transfer ratio T^{-1}. It is not claimed that this latter viewpoint will ultimately be more advantageous than the first, but only that two viewpoints are better than one.

For single-error correction in a block of length n containing b buffer positions and $k = n - b$ information positions we need only have a receiver filter with an impulse response with period of length n with each b successive digits in that response different from each other subsequence of length b. This is possible for the case $n = 2^b - 1$ and the proper polynomial is one of degree b which has a maximal-length "null sequence"(1) of $2^b - 1$ digits. Several of these are listed in Fig. 5.

IV A Multiple-Error Correcting Coding Scheme

We now consider the coding of a block of seven digits, two of which are information digits and five of which are buffer digits. We shall use for the transmitter filter one having a transfer ratio $T = (D^2 + D + I)/(D + I)$ and for the receiver filter one with the inverse ratio $T^{-1} = (D + I)/(D^2 + D + I)$. To test the properties of this set of filters in the detection and correction of errors we shall be interested only in the last five digits (in the block of seven digits) of the receiver filter response to noise. Only if a noise pattern causes a distinctive subsequence of five digits to appear in the buffer positions of the X' sequence can this noise pattern be recognized and corrected.

The impulse response of the receiver filter is given in Fig. 6-a. In Fig. 6-b are listed the seven possible responses of this filter to single noise digits (single errors). Note that the last five digits of these patterns are all different and that these noise patterns may therefore be detected and corrected.

In Fig. 6-c are listed the twenty-one possible double error patterns and the responses they produce in the receiver filter. (These may be found by adding, modulo-two, the proper single-error responses.) Three of the responses are, in their final five digits, the same as those produced by certain single errors. Therefore when one of these (starred) sub-sequences is received in the buffer positions of X' it will be interpreted as due to a single, rather than a double, error since the former is more likely than the latter.

Of the thirty-one sequences of five digits possible (excluding the all-zero combination, which is interpreted as "no error") six of them cannot occur due to single or double errors. Those

six are listed in Fig. 6-d along with the possible pairs of triple-errors which could cause them. Whenever one of these six sub-sequences is received in the buffer positions of X' two equally probable noise sequences (each containing three ones) are the possible, and the most likely, causes. Therefore in decoding it makes no difference which of these two we assume for the error pattern.

The net result of the use of the filters described above is that thirty-two sub-sequences are possible in the last five digits of the sequence X'. These correspond to no error and to the thirty-one single and multiple-error patterns which are listed in Fig. 7-a. These thirty-one sequences constitute the "sphere" which surrounds the transmitted sequence (Z) = (0000000). The other three transmitted Z sequences are given in Fig. 7-b. These were determined by scrambling the other three X sequences in the transmitter filter with transfer ratio $T = (D^2+D+I)/(D+I)$. The "spheres" surrounding each of the remaining three Z sequences can be found by adding to each such sequence the thirty-one non-zero sequences of Fig. 7-a.

The error probability associated with the coding method above is easily calculated. Let p be the error probability for a single digit in the channel, and $q = 1 - p$. Then since in the sphere surrounding the transmitted message sequence the number of points at distance one is seven, the number at distance two is eighteen, and the number at distance three is six, the error probability is

$$P_e = q^7 + 7q^6p + 18q^5p^2 + 6q^4p^3 \qquad (10)$$

Slepian[3] has shown that this is the minimum possible error probability for $k = 2$, $n = 7$.

V Summary

In Fig. 8 are listed the transfer ratios for transmitter filters which yield the minimum error probabilities listed by Slepian[3]. These expressions have been arrived at by a variety of methods and are not necessarily those which require a minimum number of shift-registers for realization. Often alternate filters exist which use the same number of shift-registers for their realization and which code for the same error-probability as those which are listed. For instance, for $k = 4$ and $n = 8$, $T = (D^2+I)/(D^4+D+I)$ may be substituted for the $T = D^4 + D^3 + I$ which is listed in the table.

The general question of finding the most economical filter for minimum possible error-probability has not yet been satisfactorily solved and the author wishes to defer presentation of his fragmentary results until such time as they cohere more satisfactorily.

It is clear that the method presented here can be extended to other than the modulo-two number system. For example, if we consider X to be a block of four ternary digits, - - - two information digits followed by two buffer digits, - - - then a transmitter filter with a transfer ratio

$$\frac{Z}{X} = D^2 + 2D + 2 \qquad (11\text{-a})$$

will encode the block for transmission in a noisy channel. Keeping in mind that the + (modulo-three addition) operation is no longer self-inverse and that $1 + 2 = 2 + 1 = 0$, it follows that

$$Z = D^2X + 2DX + 2X \qquad (11\text{-b})$$

and

$$X + 2Z + Z = X + 2Z + D^2X + 2DX + 2X \qquad (11\text{-c})$$

and

$$X = D^2X + 2DX + 2Z \qquad (11\text{-d})$$

From these expressions it may be found that the impulse response of the transmitter filter is

$$- - - \ 0000\underline{2}2100000000 \ - - - \qquad (12\text{-a})$$

and that of the receiver filter is

$$- - \ -0000\underline{2}1101220\underline{2}1101220\underline{2}1101 - - - \qquad (12\text{-b})$$

The sixteen encoded sequences, Z, in Fig. 9-a may then be derived and the noise patterns due to errors in single positions may be calculated to be those of Fig. 9-b. Note that the last two (buffer) positions can contain one of the eight possible (non-zero) combinations of two ternary digits corresponding to the eight possible kinds of single errors which may be detected and corrected. That this is possible is due to the fact that the null sequence associated with $D^2 + 2D + 2$ is of maximal length: $3^2 - 1 = 8$. In this case, again, we have chosen to concentrate our attention on the impulse response of the receiver filter rather than on the somewhat more elusive "distance" properties of the encoded messages.

In summary, the viewpoint introduced here suggests that, instead of thinking about the distance properties of 2^k message points in an n-dimensional space, we may profitably think of designing a linear binary sequence filter at the receiver whose impulse response is of such a form that, by viewing $b = n - k$ successive digits of it we distinguish sub-sequences due to single errors, by viewing b digits of two superimposed impulse responses we may distinguish sub-sequences due to double errors, etc. (Regardless of how complicated the desired impulse response may be we are certain that there exists a corresponding linear sequence filter.) The corresponding messages encoded by the transmitter filter have the property that, when they are fed into the receiver filter in its "rest" state, they leave the filter in its "rest" state at the end of the message sequence. Finally, we might propose the question: For a given n and k, and for a fixed number of shift-registers what transfer ratio should the receiver filter have to minimize the associated error probability?

References:

(1) D.A. Huffman, "The Synthesis of Linear Sequential Coding Networks", Proceedings of the Third London Symposium on Information Theory, September 13, 1955.

(2) R.W. Hamming, "Error Detecting and Error Correcting Codes", Bell System Technical Journal, pp. 147-160, 1950

(3) D. Slepian, "A Class of Binary Signaling Alphabets", Bell System Technical Journal, pp. 203-234, 1956.

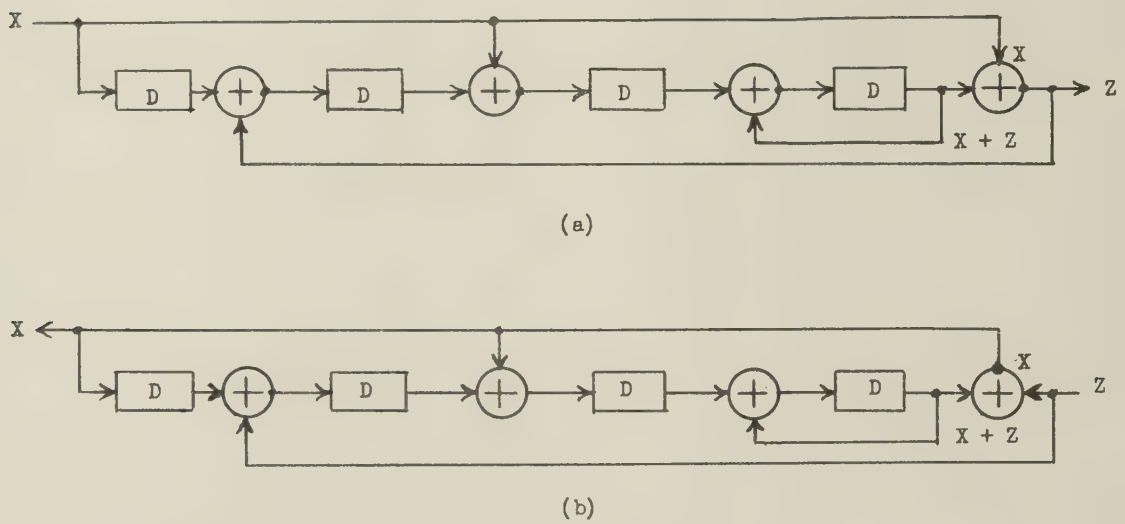


Fig. 1 - Chain realization of a binary sequence filter and its inverse.

X^*: . . . 0 0 0 . 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . . .

Z^*: . . . 0 0 0 . 1 0 1 0 1 0 0 1 1 1 0 1 0 0 1 1 1 0 1 0 0 . . .

Z_p^*: . . . 0 0 0 . 1 1 1 0 1 0 0 , 1 1 1 0 1 0 0 , 1 1 1 0 1 0 0 , . . .

Z_t^*: . . . 0 0 0 . 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . . .

(a)

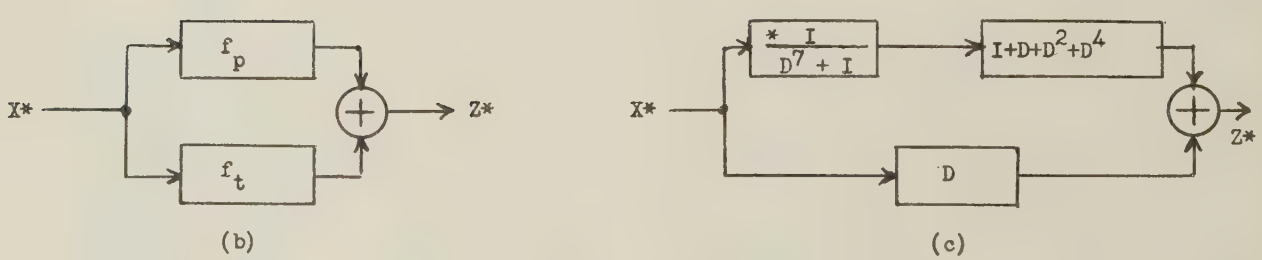
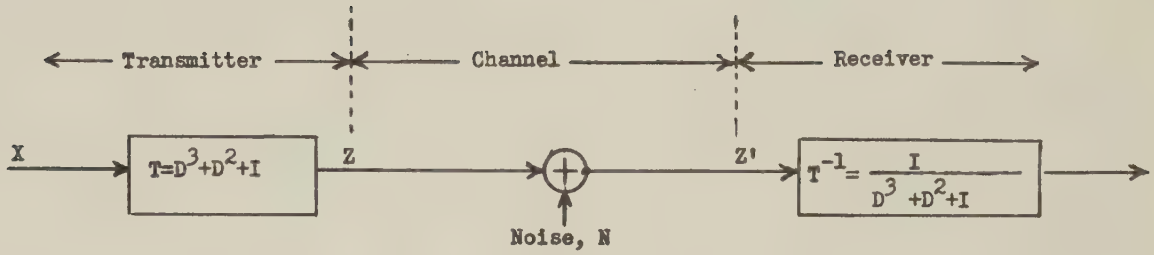


Fig. 2 - Steps in the synthesis of a binary filter from a specified impulse response.



(a)

X: (1 1 1 0) 0 0 0

Z: 1 1 0 0 0 1 0

N: 0 0 1 0 0 0 0

Z': 1 1 1 0 0 1 0

X': 0 0 1 0 1 1 1

(b)

X: (1 1 1 0) 0 0 0

(T^{-1}) N: 0 0 1 0 1 1 1

$X' = X + (T^{-1})$ N: 1 1 0 0 1 1 1

(c)

Impulse response of the receiver filter with transfer ratio $T^{-1} = (D^3 + D^2 + I)^{-1}$:

. . . 0 0 0 0 0 . 1 0 1 1 1 0 0 , 1 0 1 1 1 0 0 , 1 0 1 . . .

(d)

Fig. 3 - An elementary example of the linear single-error detecting scheme.

X:	Z = (T)X:
0 0 0 0 | 0 0 0	0 0 0 0 0 0 0
0 0 0 1 | 0 0 0	0 0 0 1 0 1 1
0 0 1 0 | 0 0 0	0 0 1 0 1 1 0
0 0 1 1 | 0 0 0	0 0 1 1 1 0 1
0 1 0 0 | 0 0 0	0 1 0 1 1 0 0
0 1 0 1 | 0 0 0	0 1 0 0 1 1 1
0 1 1 0 | 0 0 0	0 1 1 1 0 1 0
0 1 1 1 | 0 0 0	0 1 1 0 0 0 1
1 0 0 0 | 0 0 0	1 0 1 1 0 0 0
1 0 0 1 | 0 0 0	1 0 1 0 0 1 1
1 0 1 0 | 0 0 0	1 0 0 1 1 1 0
1 0 1 1 | 0 0 0	1 0 0 0 1 0 1
1 1 0 0 | 0 0 0	1 1 1 0 1 0 0
1 1 0 1 | 0 0 0	1 1 1 1 1 1 1
1 1 1 0 | 0 0 0	1 1 0 0 0 1 0
1 1 1 1 | 0 0 0	1 1 0 1 0 0 1

Fig. 4 - Coded sequences for single-error correction (n = 7).

$D^2 + D + I$	$D^5 + D^3 + I$
$D^3 + D^2 + I$	$D^6 + D^5 + I$
$D^4 + D^3 + I$	$D^7 + D^6 + I$

Fig. 5 - Polynomials having maximal length null sequences.

(a) Impulse response of filter:

. . . 0 0 0 1 0 1 1 0 1 1 0 1 1 0 1 1 . . .

(b) Response of filter to single errors:

N:		$(T^{-1})N$:	
1	0 0 0 0 0 0 0		1 0 1 1 0 1 1
	0 1 0 0 0 0 0		0 1 0 1 1 0 1
	0 0 1 0 0 0 0		0 0 1 0 1 1 0
	0 0 0 1 0 0 0		0 0 0 1 0 1 1
	0 0 0 0 1 0 0		0 0 0 0 1 0 1
	0 0 0 0 0 1 0		0 0 0 0 0 1 0
	0 0 0 0 0 0 1		0 0 0 0 0 0 1

(c) Response of filter to double errors:

N:		$(T^{-1})N$:	
1	1 0 0 0 0 0 0		1 1 1 0 1 1 0 * *
	1 0 1 0 0 0 0		1 0 0 1 1 0 1 *
	1 0 0 1 0 0 0		1 0 1 0 0 0 0
	1 0 0 0 1 0 0		1 0 1 1 1 1 0
	1 0 0 0 0 1 0		1 0 1 1 0 0 1
	1 0 0 0 0 0 1		1 0 1 1 0 1 0 *
	0 1 1 0 0 0 0		0 1 1 1 0 1 1 *
	0 1 0 1 0 0 0		0 1 0 0 1 1 0
	0 1 0 0 1 0 0		0 1 0 1 0 0 0
	0 1 0 0 0 1 0		0 1 0 1 1 1 1
	0 1 0 0 0 0 1		0 1 0 1 1 0 0
	0 0 1 1 0 0 0		0 0 1 1 1 0 1
	0 0 1 0 1 0 0		0 0 1 0 0 1 1
	0 0 1 0 0 1 0		0 0 1 0 1 0 0
	0 0 1 0 0 0 1		0 0 1 0 1 1 1
	0 0 0 1 1 0 0		0 0 0 1 1 1 0
	0 0 0 1 0 1 0		0 0 0 1 0 0 1
	0 0 0 1 0 0 1		0 0 0 1 0 1 0
	0 0 0 0 1 1 0		0 0 0 0 1 1 1
	0 0 0 0 1 0 1		0 0 0 0 1 0 0
	0 0 0 0 0 1 1		0 0 0 0 0 1 1

(d) Response of filter to certain triple errors:

N:		$(T^{-1})N$:	
1	0 0 1 0 1 1 0		0 0 1 0 0 0 1
	1 0 0 1 0 0 1		1 0 1 0 0 0 1
	0 0 1 0 1 0 1		0 0 1 0 0 1 0
	1 0 0 1 0 1 0		1 0 1 0 0 1 0
	0 0 1 0 0 1 1		0 0 1 0 1 0 1
	1 0 0 1 1 0 0		1 0 1 0 1 0 1
	0 0 1 1 1 0 0		0 0 1 1 0 0 0
	1 0 0 0 0 1 1		1 0 1 1 0 0 0
	0 0 1 1 0 0 1		0 0 1 1 1 0 0
	1 0 0 0 1 1 0		1 0 1 1 1 0 0
	0 0 1 1 0 1 0		0 0 1 1 1 1 1
	1 0 0 0 1 0 1		1 0 1 1 1 1 1

Fig. 6 - Error response of $(D + I)/(D^2 + D + I)$.

(0 0 0 0 0 0 0)	1 0 0 1 0 0 0	0 0 1 1 0 0 0	0 0 0 0 1 0 1	
1 0 0 0 0 0 0	1 0 0 0 1 0 0	0 0 1 0 1 0 0	0 0 0 0 0 1 1	
0 1 0 0 0 0 0	1 0 0 0 0 1 0	0 0 1 0 0 1 0	0 0 1 0 1 1 0	(or 1001001)
0 0 1 0 0 0 0	1 0 0 0 0 0 1	0 0 1 0 0 0 1	0 0 1 0 1 0 1	(or 1001010)
0 0 0 1 0 0 0	0 1 0 1 0 0 0	0 0 0 1 1 0 0	0 0 1 0 0 1 1	(or 1001100)
0 0 0 0 1 0 0	0 1 0 0 1 0 0	0 0 0 1 0 1 0	0 0 1 1 1 0 0	(or 1000011)
0 0 0 0 0 1 0	0 1 0 0 0 1 0	0 0 0 1 0 0 1	0 0 1 1 0 0 1	(or 1000110)
0 0 0 0 0 0 1	0 1 0 0 0 0 1	0 0 0 0 1 1 0	0 0 1 1 0 1 0	(or 1000101)

(a) Shape of "sphere" surrounding transmitted sequences

X:			Z:
0 0 0 0 0 0 0	⟶		0 0 0 0 0 0 0
0 1 0 0 0 0 0			0 1 0 1 1 1 1
1 0 0 0 0 0 0			1 0 1 1 1 1 1
1 1 0 0 0 0 0			1 1 1 0 0 0 0

$$\frac{Z}{X} = \frac{D^2 + D + I}{D + I}$$

(b) Derivation of the four transmitted sequences

Fig. 7 - Results of coding with binary filter for k = 2, n = 7.

For k = 2:

n = 4,	$D^2 + D + I$
n = 5,	$(D^2 + D + I)/(D + I)$
n = 6,	$I/(D^2 + I)$
n = 7,	$(D^2 + D + I)/(D + I)$
n = 8,	$I/(D^2 + D + I)$
n = 9,	$(D^3 + D + I)/(D^2 + I)$
n = 10,	$I/(D^2 + D + I)$
n = 11,	$I/(D^2 + D + I)$
n = 12,	$(D^3 + D + I)/(D^2 + I)$

For k = 3:

n = 5,	$D^2 + D + I$
n = 6,	$D^3 + D^2 + I$
n = 7,	$D^4 + D^3 + I$
n = 8,	$(D^2 + I)/(D^4 + D^3 + I)$
n = 9,	$I/(D^3 + D + I)$
n = 10,	$I/(D^3 + D + I)$
n = 12,	$I/(D^3 + D + I)$

For k = 4:

n = 6,	$D^2 + D + I$
n = 7,	$D^3 + D^2 + I$
n = 8,	$D^4 + D^3 + I$

For k = 5:

n = 7,	$D^2 + D + I$
n = 8,	$D^3 + D^2 + I$
n = 9,	$D^4 + D^3 + I$

For k = 6:

n = 8,	$D^2 + D + I$
n = 9,	$D^3 + D^2 + I$
n = 10,	$D^4 + D^3 + I$
n = 11,	$D^5 + D^3 + I$

For k = 7:

n = 9,	$D^2 + D + I$
n = 10,	$D^3 + D^2 + I$
n = 11,	$D^4 + D^3 + I$
n = 12,	$D^5 + D^3 + I$

For k = 8:

n = 10,	$D^2 + D + I$
n = 11,	$D^3 + D^2 + I$
n = 12,	$D^4 + D^3 + I$

Fig. 8 - Transmitter filters for minimum error-probability.

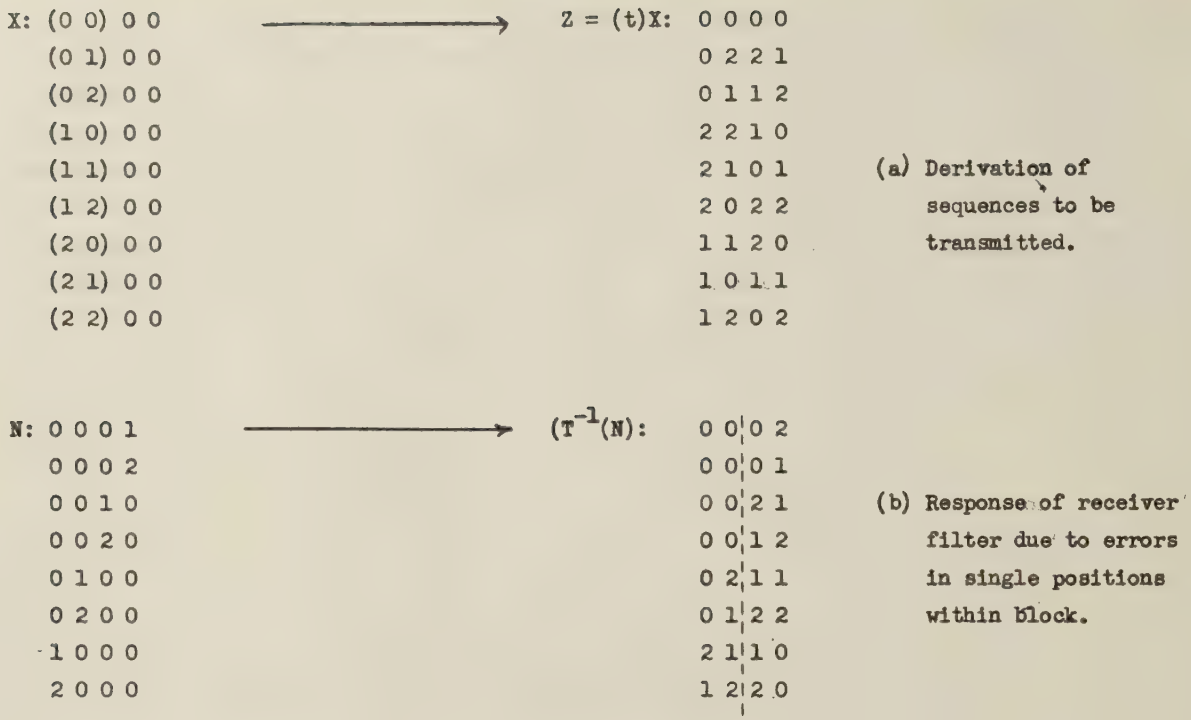


Fig. 9 - Encoded sequences and error patterns for a ternary linear coding scheme; $T = D^2 + 2D + 2$.

THEORY OF INFORMATION FEEDBACK SYSTEMS

Sheldon S. L. Chang
New York University
University Heights
New York 53, N. Y.

Abstract

A general information feedback system is defined and formulated in a way broad enough to allow coded or uncoded channels with total or partial information feedback. Basic theorems governing change in information rate and reliability are derived with full consideration of the transition probabilities of both direct and feedback channels, including message words as well as the confirmation - denial signal.

The process involved in this type of feedback is as follows. There are a set of $\{x\}$ words which are transmitted and a set of $\{y\}$ words received. The set $\{X\}$ is divided, as previously agreed upon into $\{X\}$ groups, such that x_{ij} is the j^{th} member of the i^{th} group. Corresponding to the transmission of x_{ij} the word $y_{i'j'k'}$ may be received. The reception of this word indicates that $x_{i'j'} \in X_{i'}$ was transmitted. The receiver therefore sends back to the transmitter the message $Y_{i'}$ to indicate the belief that a message from the group $X_{i'}$ was sent. Due to noise in the feedback channel the message $Z_{i''}$ is received. If $Z_{i''}$ corresponds to X_i, the sender confirms the report by sending the subsequent x. If $Z_{i''}$ does not correspond to X_i, the sender transmits the denial signal. It is assumed that while the feedback channel is reporting, the direct channel is sending new information simultaneously. An example is given which shows how this is accomplished with delayed feedback.

The following theorem and corollary are derived:

Theorem I. The gain in information at the receiver by the confirmation - denial process is less than the average signal entropy required for the confirmation - denial transmission by

$$\sum_{ii'} P(X_i Y_{i'}) H_{ii'}$$

where $P(X_i, Y_{i'})$ is the joint probability that a word from the X_i group is transmitted and a word from the $Y_{i'}$ group is received, and

$$H_{ii'} = - P_{ii'} \log P_{ii'} - (1-P_{ii'}) \log (1-P_{ii'})$$

is the entropy associated with the probability $P_{ii'}$ of the feedback channel. This is the probability that when a word from the X_i group is sent and a word from the $Y_{i'}$ group is received, the receiver will obtain confirmation through the feedback process. $H_{ii'}$ is therefore the uncertainty of confirmation or denial when a word from the X_i group is sent and a word from the $Y_{i'}$ group is received.

Corollary I. The gain in information at the receiver due to the reduction in equivocation by the confirmation - denial process is equal to the net information of the confirmation - denial signal when the feedback channel is error-free.

A special case of information feedback systems is the discarding system in which a wrong message is corrected by erasing and then repeating. If the feedback message indicates that an erasure was incorrectly received as a message word, the sender transmits two erasures to erase both the message word and the preceding incorrect message. When the feedback message indicates that a message has been incorrectly received as an erasure, the sender transmits, not another erasure, but simply the preceding message and the intended message.

In a discarding system there is a natural iterative process which greatly reduces the error probabilities of the confirmation - denial signal. The information that a particular message is confirmed or denied is not entirely contained in the absence or presence of an erasure signal. Each succeeding message is a confirmation of a previous choice.

Theorem II. In a iterative discarding system, the error probabilities of the confirmation-denial signal are successively reduced by the iterative process. If the feedback channel is error free, and if the feedback group containing the erasure signal contains it as the one and only signal, the error probabilities decay to zero at approximately an exponential rate with each subsequent confirmed signal. Otherwise they converge rapidly to constants.

A third theorem deals with the equivocation of the finally accepted signal.

Some deductions from the basic theorems are listed below:

(1) With error-free feedback of sufficient capacity to repeat the whole message as it is transmitted, information can be sent error-free through a noisy channel with little or no loss of the net information rate and without coding.

(2) With an equally noisy feedback channel, feedback is by far more effective than coding in error reduction. By effective is meant that there is no need of elaborate coding nor serious loss of channel capacity.

Classification of Feedback Systems

The utilization of feedback in a two way communication system to improve message reliability is a natural process which occurs at least millions of times daily. During a telephone conversation, while A is talking to B, B may ask A to repeat upon hearing an indistinguishable word, or A may ask B to repeat if the message is important or without redundancy and must be reliably transmitted. Aside from feedback relating to the immediate word or sentence, A may talk at a slower rate if he is asked to repeat too often or if B makes mistakes too frequently in recital. Alternately B may ask A to talk slower if what he hears is not clear enough.

The above natural processes are not efficient in terms of information rate or improvement of reliability. However, they embody the essential principles upon which efficient feedback communication systems may be devised.

There are two distinct functions of the feedback process, as illustrated above:

(i) improvement of the immediate message

(ii) matching of the communication rate to the noise level.

To devise means for fulfilling the second function is essentially an instrumentation problem. Its very existence depends on the assumption that the signal to noise ratio is a slowly varying function. Consequently, it is allowable to assume that within any given interval, which may contain many words duration, both the signal to noise ratio and the communication rate do not change appreciably. Once the basic laws governing the information rate and reliability under constant signal to noise ratio and communication rate are established, it is always possible to design a system which automatically adjusts for the optimum condition.

As a preliminary study this paper will be concerned with the more basic problem, that of improvement of the immediate message by means of feedback, while signal to noise ratio and communication rate are held constant.

Generally speaking, there are two types of feedback systems which improve the immediate message:

(1) decision feedback

(2) information feedback.

In case (1), the sender adds redundancy or coding into the message so that out of a total of N possible sequences, only M are selected as alphabets for conveying information and are transmitted. Of the remaining N-M sequences, there may be M' sequences such that each of the M' is close enough to one of the M and far away from the other M-1 in signal space that it can be interpreted as the former without much risk of being in error. Upon reception of one of M or M', the receiver will record it and report "+" which means "please proceed". If one of the N-(M+M') is received, the receiver will not record and report "-" which means "please repeat", and the transmitter will then repeat the information.

There is no need for the transmitter to wait while receiving the report from the receiver. It may transmit subsequent information at the same time if, for instance, the agreed arrangement is that at a "-" report the next two sequences are subsequent informations while the third is the requested repetition, or similar arrangements.

For this system to function well, the feedback channel must be able to transmit reliably one bit of information during the interval that the direct channel transmits one sequence. Alternatively, the feedback channel may be required to report the positions of received sequences belonging to the N-(M+M') after each group of sequences instead of reporting "+" or "-" after each sequence and the required feedback channel capacity would be even smaller.

The essential problem in this type of system is to code the direct transmission channel into codes which allow for nulls or rejected information.* The feedback is simply a means of filling the null positions with additional information from the direct transmission channel. If the feedback channel has negligible error, which is not difficult to obtain as the required channel capacity is low, the information rate and the probability of error of the direct transmission channel will remain unaltered.

Hereafter case (2), information feedback, will be considered.

I. Information Feedback, Definition, General Theorem

In an information feedback system, the receiver reports back in whole or in part the received information and the sender will decide whether or not he is satisfied with the information as received, and in the latter event, he will send corrective information.

In calculating channel capacity, it is assumed as in case (1) that while the feedback channel is reporting, the direct transmission channel is sending new information simultaneously so that no standby period will be allowed for,**

The basic process involved in this type of

* For instance, Hammings[3] double error detecting single error correcting code, and others as given in reference 2.

** This point will be illustrated later in section II.

feedback is as follows: There are a set of $\{x\}$ words which are transmitted and a set of $\{y\}$ words received. The set of $\{x\}$ is divided, as previously agreed upon into $\{X\}$ groups, such that x_{ij} is the jth member of the i^{th} group. Corresponding to the transmission of x_{ij} the word $y_{i'j'k'}$ may be received. The reception of this word indicates that $x_{i'j'} \in X_{i'}$ was transmitted. The receiver therefore sends back to the transmitter the message $Y_{i'}$ to indicate the belief that a message from the group $x_{i'}$ was sent. Due to noise in the feedback channel the message $Z_{i''}$ is received. If $Z_{i''}$ corresponds to X_i, the sender confirms the report. If $Z_{i''}$ does not correspond to X_i, the sender transmits the denial signal.

The above processes have the following physical significance.

1. There is an improvement in reliability, since only confirmed messages are finally retained.

2. There are two opposite changes in information rate. The sender has to provide additional capacity to keep the receiver informed as to whether the reported message is correct.

However, there is additional information gained by the receiver in knowing whether or not the original message is in the reported group.

3. The feedback channel may have considerably lower capacity than the direct transmission channel since each group may contain many words, yet the improvement in reliability can still be substantially realized. It is generally possible to assign the transmitted words in such a way that words in the same group are far away from each other in the signal space. Once confirmed, there is little probability of error, as the other words in the same group are very unlikely to be received by mistake.

In the following, a quantitative study will be made on the above mentioned effects.

After receiving $y_{i'j'k'}$, the posteriori probability that the transmitted word was x_{ij} is $P_{y_{i'j'k'}}(x_{ij})$. After the receiver reports $Y_{i'}$, the conditional probability that the sender will receive $Z_{i''}$ is $P_{Y_{i'}}(Z_{i''})$. When $Z_{i''}$ corresponds to X_i, and the sender confirms the report, there is a probability $P_c(d)$ that the confirmation is received as a denial. Similarly, when $Z_{i''}$ does not correspond to X_i, there is a probability $P_d(c)$ that the denial is taken as a confirmation. There are, of course, the probabilities $P_c(c)$ and $P_d(d)$ that the confirmation and denial signals are correctly received. From these definitions

$$\left.\begin{aligned} P_c(c)+P_c(d) &= 1 \\ P_d(c)+P_d(d) &= 1 \end{aligned}\right\} \qquad (1)$$

and $$\sum_{i''} P_{Y_{i'}}(Z_{i''}) = 1 \qquad (2)$$

There are various ways of transmitting the confirmation denial signal. One method of special interest is for the sender to transmit the denial signal only. When he transmits the subsequent message, confirmation of the previous report from the receiver is automatically implied. While this particular method will be discussed in detail in subsequent sections, the discussion here is kept at a general level, without any reference as to how the confirmation denial information is conveyed.

Upon receiving $y_{i'j'k'}$, the probability that x_{ij} was sent and a report of $Y_{i'}$ will be confirmed is

$$P_{y_{i'j'k'}}(x_{ij})\, P_{ii'} \qquad (3)$$

where $P_{ii'}$ is defined as

$$P_{ii'} = P_{Y_{i'}}(Z_i)P_c(c) + [1-P_{Y_{i'}}(Z_i)]P_d(c) \qquad (4)$$

and the feedback messages Z_i correspond to X_i. $P_{ii'}$ is the conditional probability that the feedback word $Y_{i'}$ will result in confirmation, when $x_{ij} \in X_i$ was originally transmitted.

Upon receiving $y_{i'j'k'}$ the probability that x_{ij} was sent and a report of $Y_{i'}$ will be denied is:

$$P_{y_{i'j'k'}}(x_{ij})[1-P_{ii'}] \qquad (5)$$

The total probability of confirmation upon receiving $y_{i'j'k'}$ is therefore

$$P_{y_{i'j'k'}}(c) = \sum_i \sum_j P_{ii'} P_{y_{i'j'k'}}(x_{ij}) \qquad (6)$$

while the total probability of denial is,

$$P_{y_{i'j'k'}}(d) = \sum_i \sum_j [1-P_{ii'}] P_{y_{i'j'k'}}(x_{ij}) \qquad (7)$$

The conditional probability, upon receiving both $y_{i'j'k'}$ and the confirmation of $Y_{i'}$, that x_{ij} was originally sent is

$$P_c\, y_{i'j'k'}\, (x_{ij}) = \frac{P_{ii'}\, P_{y_{i'j'k'}}(x_{ij})}{P_{y_{i'j'k'}}(c)} \qquad (8)$$

and similarly, the conditional probability upon receiving both $y_{i'j'k'}$ and the denial of $Y_{i'}$, that x_{ij} was sent is

$$P_d\, y_{i'j'k'}\, (x_{ij}) = \frac{(1-P_{ii'})P_{y_{i'j'k'}}(x_{ij})}{P_{y_{i'j'k'}}(d)} \qquad (9)$$

The equivocation after receiving both $y_{i'j'k'}$ and subsequent confirmation is therefore:

$$H_{cy_{i'j'k'}}(x) = -\sum_i \sum_j P_{cy_{i'j'k'}}(x_{ij}) \times$$

$$\log P_{cy_{i'j'k'}}(x_{ij}) \qquad (10)$$

By substitution and performing the summation where possible,

$$H_{cy_{i'j'k'}}(x) = \log P_{y_{i'j'k'}}(c) - \sum_i \sum_j P_{cy_{i'j'k'}}(x_{ij})$$

$$\left\{ \log P_{ii'} + \log P_{y_{i'j'k'}}(x_{ij}) \right\} \qquad (11)$$

Similarly,

$$H_{dy_{i'j'k'}}(x) = \log P_{y_{i'j'k'}}(d) -$$

$$- \sum_i \sum_j P_{dy_{i'j'k'}}(x_{ij}) \times$$

$$\left\{ \log (1-P_{ii'}) + \log P_{y_{i'j'k'}}(x_{ij}) \right\} \qquad (12)$$

On the average, the total equivocation is reduced by the confirmation - denial process by the amount

$$\Delta I_{y_{i'j'k'}} = H_{y_{i'j'k'}} - P_{y_{i'j'k'}}(c) H_{cy_{i'j'k'}}(x) -$$

$$- P_{y_{i'j'k'}}(d)\, H_{dy_{i'j'k'}}(x) \qquad (13)$$

whence, by substitution, and noting that by definition the equivocation is,

$$H_{y_{i'j'k'}} = -\sum_i \sum_j P_{y_{i'j'k'}}(x_{ij}) \times$$

$$\log P_{y_{i'j'k'}}(x_{ij}) \qquad (14)$$

the result is obtained that

$$\Delta I_{y_{i'j'k'}} = -P_{y_{i'j'k'}}(c) \log P_{y_{i'j'k'}}(c)$$

$$- P_{y_{i'j'k'}}(d) \log P_{y_{i'j'k'}}(d)$$

$$+ \sum_i \sum_j P_{y_{i'j'k'}}(x_{ij}) \left\{ P_{ii} \log P_{ii} \right.$$

$$\left. + (1-P_{ii'}) \log (1-P_{ii'}) \right\} \qquad (15)$$

The signal entropy for the confirmation - denial process is

$$H_{y_{i'j'k'}}(cd) = -P_{y_{i'j'k'}}(c) \log P_{y_{i'j'k'}}(c)$$

$$- P_{y_{i'j'k'}}(d) \log P_{y_{i'j'k'}}(d) \qquad (16)$$

while an uncertainty entropy $H_{ii'}$ may be defined as,

$$H_{ii'} = -P_{ii} \log P_{ii'} - (1-P_{ii'}) \log (1-P_{ii'}) \qquad (17)$$

The reduction in the total entropy of the received message by the confirmation denial process is then concisely given as

$$\Delta I_{y_{i'j'k'}} = H_{y_{i'j'k'}}(cd)$$

$$- \sum_i \sum_j P_{y_{i'j'k'}}(x_{ij})\, H_{ii'} \qquad (18)$$

This relation shows that the average gain in information by the confirmation - denial process is less than the entropy of the code required for the confirmation - denial process. The net difference is

$$H_{y_{i'j'k'}}(cd) - \Delta I_{y_{i'j'k'}}$$

$$- \sum_i \sum_j P_{y_{i'j'k'}}(x_{ij})\, H_{ii'} \qquad (19)$$

The symbol $\sum_{i'j'k'}$ will indicate a summation or an integration over the received signals, according to whether the received signals are discrete or continuous. The probability that the received signal is $y_{i'j'k'}$ is $P(y_{i'j'k'})$ if y is discrete or $p(y_{i'j'k'})\, dK'$ if y is continuous. Hence the difference average over all $y_{i'j'k'}$ is

$$H(cd) - \Delta I = \qquad (20)$$

$$- \sum_{i'j'k'} \sum_{i\,j} P(y_{i'j'k'}) P_{y_{i'j'k'}}(x_{ij}) H_{ii'}$$

where $H_{(cd)}$ is the averaged entropy of the code required for confirmation - denial. By combining

$$H(cd) - \Delta I$$

$$= \sum_{i'j'k'} \sum_{ij} P(x_{ij}, y_{i'j'k'})\, H_{ii'} \qquad (21)$$

Where $P(y_{i'j'k'}, x_{ij})$ is the joint probability that x_{ij} is sent and $y_{i'j'k'}$ is received. Performing the summation over j',k', and j

$$\sum_{j'k'j} P(x_{ij}, y_{i'j'k'}) = P(X_i, Y_{i'}) \qquad (22)$$

is the joint probability that a word from the X_i group is sent and a word from the Y_i group is received.

Thus the net difference is

$$\sum_{ii'} P(X_i, Y_{i'}) H_{ii'} \qquad (23)$$

Restated as a theorem the results are:

Theorem I The gain in information at the receiver by the confirmation denial process is less than the average signal entropy required for the confirmation - denial transmission by

$$\sum_{ij'} P(X_i Y_{i'}) H_{ii'}$$

where $P(X_i, Y_{i'})$ is the joint probability that a word from the X_i group is transmitted and a word from the Y_i group is received, and

$$H_{ii'} = - P_{ii'} \log P_{ii'} - (1-P_{ii'}) \log (1-P_{ii'})$$

is the entropy associated with the probability $P_{ii'}$ of the feedback channel. This is the probability that when a word from the X_i group is sent and a word from the $Y_{i'}$ group is received, the receiver will obtain confirmation through the feedback process. $H_{ii'}$ is therefore the uncertaintly of confirmation or denial when a word from the X_i group is sent and a word from the Y_i group is received.

Two Special Cases An interesting case is when the feedback channel is error-free. Formally,

$$P_{Y_{i'}}(Z_i) = \delta_{ii'}$$

and, $P_{ii'} = \delta_{ii'} P_c(c) + (1- \delta_{ii'}) P_d(c)$.

Then the uncertainty entropy for $i = i'$ becomes,

$$H_{ii} = H_{i'i'} = - P_c(c) \log P_c(c) - [1-P_c(c)] \log [1-P_c(c)] \qquad (24)$$

and for $i \neq i'$

$$H_{ii'} = - P_d(d) \log P_d(d) - [1-P_d(d)] \log [1-P_d(d)] \qquad (25)$$

The difference between the entropy of the confirmation - denial code and the gain in information at the receiver is then

$$(H(cd) - \Delta I) = \sum_{i} P(X_i, Y_i) H_{ii} + \sum_{i \neq i'} P(X_i, Y_{i'}) H_{ii'} \qquad (26)$$

But $P(X_i y_i)$ is the probability, when the feedback channel is error-free, that a confirmation will be subsequently transmitted; while $P(X_i, y_{i'})$ is the probability of a denial being sent.
Hence $\sum P(X_i, Y_i) = P(c)$

$$\sum_{i \neq i'} P(X_i, Y_{i'}) = P(d) \qquad (27)$$

Hence the difference between H(cd) and ΔI is the equivocation of the direct channel to the transmission of the confirmation - denial code.

Corollary I The gain in information at the receiver due to the reduction in equivocation by the confirmation - denial process is equal to the net information of the confirmation - denial signal when the feedback channel is error-free.

Another interesting case is that the confirmation - denial process may be assumed error-free. This is expressed mathematically as:

$$P_d(c) = P_c(d) = 0$$

$$P_d(d) = P_c(c) = 1$$

and $P_{ii'}$ reduces to the simple relation

$$P_{ii} = P_{Y_{i'}}(Z_i) \qquad (28)$$

Stated as a corollary, this result is:

Corollary II Even with an error-free confirmation - denial process, noise in the feedback channel results in a net loss in the direct channel information rate. The loss is given by

$$\text{Loss} = \sum_{ii'} P(X_i, Y_{i'}) H_{ii'}$$

where

$$H_{ii'} = - P_{y_{i'}}(Z_i) \log P_{y_{i'}}(Z_i) - [1-P_{y_{i'}}(Z_i)] \log [1-P_{y_{i'}}(Z_i)]$$

is the equivocation of the feedback channel.

Approximate Equations for Change in Information Rate

While $H_{ii'}$ can be evaluated for special systems, its evaluation in general is by no means simple. Upper bounds for the difference between the entropy of confirmation - denial signal and the gain in information may be more easily determined. This difference is greatest for the largest $H_{ii'}$. Hence, from equation 23,

$$[H(cd) - \Delta I] < \sum_{ii'} P(X_i, Y_i) H_{ii'}(\max)$$

or,

$$[H(cd) - \Delta I] < H_{ii'}(\max) \qquad (29)$$

since,

$$\sum_{ii'} P(X_i Y_{i'}) = 1$$

A lower upper bound may be determined by the following consideration. For the feedback to be at all effective the probability of

obtaining confirmation, when a word from the X_i group is sent and a word from the corresponding Y_i group is received, must be better than one - half.

$$P_{ii} > 1/2$$

or correspondingly

$$P_{ii'} < 1/2 \text{ when } i \neq i'$$

Therefore the maximum value of $H_{ii'}$ occurs for some i=i'. A more refined upper bound for the loss in information is thus

$$\left.\begin{aligned} \text{Loss} &< \sum_i P(X_i)\, H_{ii} \\ \text{or}\quad \text{Loss} &< \sum_{i'} P(Y_{i'})\, H_{i'i'} \end{aligned}\right\} \qquad (30)$$

where $P(X_i)$ is the probability of transmitting a word from the X_i group, and $P(Y_{i'})$ is the probability of receiving a word in the $Y_{i'}$ group.

The relation that gives the least upper bound should be used.

II Discarding Systems

A discarding system is a special case of information feedback systems. It is characterized by the following:

1. The denial signal or erasure, $\emptyset$ is transmitted as one of $\{x\}$, and received as the corresponding member (or members) of $\{y\}$.

2. Confirmation is implied with the transmission of subsequent information.

3. A denied word is corrected by repeating the same word following a denial. The remaining information in a denied word is totally discarded.

General Relation for the Signal Entropy of the Confirmation - Denial Process.

Consider a code where the received signals may be any one of the states y_s. If the probability of receiving y_s is $P(y_s)$ then

$$\sum_s p(y_s) = 1$$

Let $\emptyset$ be the denial or alerting signal. If the probability of receiving $\emptyset$ is $P(\emptyset)$ then the probabilities $P(y_s)$ must accordingly be reduced in the presence of the confirmation - denial process to

$$[1-P(\emptyset)]\, P(y_s)$$

The required signal entropy is then

$$-P(\emptyset)\log P(\emptyset) - \sum_s [1-P(\emptyset)]P(y_s)\log[1-P(\emptyset)]P(y_s)$$

or

$$-P(\emptyset)\log P(\emptyset)) - [1-P(\emptyset)]\log[1-P(\emptyset)]$$

$$-[1-P(\emptyset)] \sum_s P(y_s)\log P(y_s)$$

The last term in this expression represents the required signal entropy for the transmission of new information when the preceding message is confirmed. The first two terms therefore represent the entropy required for the confirmation - denial process. Since $P(\emptyset)$ represents the probability of confirmation P(c) and $[1 - P(\emptyset)]$ represents the probability of denial P(d), the entropy required for the confirmation-denial process is

$$-P(c)\log P(c) - P(d)\log P(d).$$

Pair Creation and Pair Annihilation

An undesirable by product of the discarding system is the non-conservation of the number of message words. A pair of message words is created by the erroneous transmission of an erasure into a message word. Similarly, a pair of message words is annihilated by the erroneous transmission of a message word into an erasure.

For transmission of information which has intrinsic redundancy, such as English language, the non-conservation effect is not serious. In the received message, the extraneous pair and the erroneous erasure are easily detectable. However, for messages without redundancy, this effect will cause dislocation of subsequent words, which may be considered totally erroneous in certain applications. In these cases, the transmitted message should be coded, and feedback may be used for further improvement of reliability.

From the probability point of view, the amount of information contained in a received word word is still uniquely defined irrespective of the problem of non-conservation. Its equivocation can be calculated from either equation 10 or equation 14 depending on whether the word is confirmed or unconfirmed. Its uncertainty of being the erasure is the partial equivocation pertaining to the confirmation - denial information. For the entire message, the probabilities of its being a certain probable transmitted message are different before and after the transmission. The amount of transmitted information is uniquely defined even though the received message may not have exactly the same length of the transmitted message.

However, for counting the number of errors in a received message, some artificial rule is necessary. Since a created pair or a annihilated pair is caused by one undetected error in the received message, it will be accounted for as such.

Iterative Discarding System

In a iterative[4] discarding system, the feedback process is used to improve the confirmation-denial information as well as the main message. Its rules are as follows: If the feedback message indicates that an incorrect direct channel message was received, the sender transmits an erasure followed by the correct message. If the feedback message indicates that an erasure was incorrectly received as a message word the sender transmits two erasures to erase both the message word and the preceding incorrect message. When the feedback message indicates that a message word has been incorrectly received as an erasure, the sender transmits not another erasure, but simply the preceding message and the intended message.

These rules may be summerized in tabular form. In this table 0, 1 represent the message symbols, Ø represents the erasure symbol, and P represents the preceding unerased symbol.

T \ R	0	1	Ø
0	-	Ø 0	P0
1	Ø1	-	P1
Ø	ØØ	ØØ	-

R=Received

T=Transmitted

The information that a particular message is confirmed or denied is not entirely contained in the absence or presence of an erasure signal. Each succeeding message is a confirmation of a previous choice. To illustrate this point, some examples of discarding systems with error-free and totally distinctive feedback channels will be given below:

Example I

As an example consider the message to be transmitted

"GONE WITH THE WIND"

Let [] denote a space. The messages are:

```
                1 2 3 4 5 6 7 8 9 1011 121314151617
Transmitted     G O N Ø N E Ø Ø Ø E [] W I W I T H...

Reported        G O P Ø N O T Ø Ø E [] W Ø W I T H...
```

The erasing procedure, indicated by the arrows, permits the correct message to be determined.

GONE WITH.....

The important result is that independent of where the errors originate, as long as the feedback is error free and totally distinctive, all errors will eventually be corrected and the final message will be error free.

One of the difficulties of this and other feedback systems is the delay resulting from the spatial separation between the sender and the receiver. A method of overcoming this difficulty, without standby operation, is to delay the correspondence between the corrective messages and the original message.

Example II

As an example assume that the sender will receive the feedback message before he is about to send the fourth succeeding letter. For convenience, arrange the message as follows:

```
G O N E
[] W I T
H [] T H
E [] W I
N D []
```

and send the message by scanning each row in turn. Each column will then do its own correcting.

Transmitted Message	Reported Message
G O N E	G O (P) E
[] W Ø T	(Ø) W Ø (U)
G [] N Ø	G [] N (F)
[] [] I Ø	[] (Q) I Ø
H Ø T Ø	H Ø T Ø
E [] W T	E (Ø) W T
N [] - H	N [] - H
- [] - I	- [] - I
- D -	- D -

The circles indicate errors in reception. Independently applying the erasure rules, as indicated by the arrows, the message is recovered.

```
G O N E
[]W I T
H []T H
E []W I
N D -
```

In this example the message was very short. If, for instance, the first chaper of Gone With the Wind was transmitted by this process the number of errors in each column would be almost equal. The time required to transmit the longest column would then be virtually equal to the time required to transmit the shortest.

For instance, in Example I, upon receiving the ninth message the receiver is not absolutely sure that an erasure was sent. However, if some other message was wrongly received as an erasure, the transmitter would have sent an O as the tenth message instead of an E. Therefore upon receiving the tenth message the receiver becomes more certain that the ninth message was correct. Of course, the tenth message could have been an O which was wrongly received as an E, but in that case, upon reporting E the eleventh message would have been $\emptyset$ instead of []. In a similar manner, later messages confirm or deny the correctness of the previous messages. By the time a long message has been transmitted, all previous confirmation-denials are error free. It is therefore not necessary to elaborately code the erasure message to obtain error free confirmation - denial.

Mathematically, the probability of an erroneous confirmation after receiving n subsequent message words is

$$P_{y_c}(x_d) = \prod_{s=1}^{n} P_{y_s}(x\emptyset) \qquad (31)$$

where $P_{y_s}(x\emptyset)$ is the conditional probability after receiving y_s, that the erasure was sent. Similarly, the probability of an erroneous denial after receiving n subsequent message words following an erasure is:

$$P_{y_d}(x_c) = P_{y_\emptyset}(x)\, P_{y_1}(x_o) \prod_{s=2}^{n} P_{y_s}(x\emptyset) \qquad (32)$$

where $P_{y_\emptyset}(x)$ is the conditional probability after receiving the erasure that a message word was actually sent and $P_{y_1}(x_o)$ is the conditional probability after receiving first subsequent message word that the erased message word was actually sent.

Since $P_{y_s}(x\emptyset)$ is much smaller than unity, upper and lower bounds u and $\underline{1}$* may be defined such that

$$u \geq -\log_e P_{y_s}(x\emptyset) \geq \underline{1} \qquad (33)$$

for all $s \neq \emptyset$. Equations 31 and 32 may be written as:

$$e^{-nu} \leq P_{y_c}(x_d) \leq e^{-n\underline{1}} \qquad (34)$$

$$P_{y_\emptyset}(x) P_{y_1}(x_o) e^{-(n-1)u} \leq P_{y_d}(x_c) \leq P_{y_\emptyset}(x) P_{y_1}(x_o) e^{-(n-1)\underline{1}}$$

*The letter "1" is underlined to distinguish it from the number "1".

The above shows that with a totally distinctive noiseless feedback channel, the probabilities of erroneous confirmation and erroneous confirmation and erroneous denial are reduced to zero at approximately an exponential rate. It is not necessary, however, for the feedback channel to be totally distinctive on the message words themselves. The above deduction is equally valid as long as the member of $\{X\}$ containing $\emptyset$ consists of $\emptyset$ only. In case y_1 and x_o happen to belong to the same group, equation 32 may appear to be wrong at first glance. However, in such a case, as a message word of the same group is finally received, the erroneous first denial is considered as having been corrected for, and the final confirmation denial signal is not in error in itself.

A more complicated situation exists if the member of $\{X\}$ containing the erasure contains message words as well and the feedback channel is not error free. In order to utilize equation 8 to determine the conditional probabilities of a finally accepted message word, an approximate expression will be derived for $P_d(c)$ which is the error probability that a word is finally confirmed while an erasure was originally transmitted. Neglecting intersymbol influence, one has approximately

$$P_d(c) = \sum_{ijk}{}' P_{x_\emptyset}(y_{ijk})\, P_{Y_i}(Z_\emptyset)\, P_c(c)$$

$$+ \sum_{ijk}{}' P_{x_\emptyset}(y_{ijk})\,[1-P_{Y_i}(Z_\emptyset)]P_d(c)$$

$$+ \sum_{ijk}{}' P_{x_\emptyset}(y_{ijk})[1-P_{Y_i}(Z_\emptyset)]P_{x_\emptyset}(y\emptyset)P_d(c) \qquad (35)$$

In equation 35, Σ' means summing over all values of $j \neq \emptyset$. On the right hand side of equation 35, the first term represents the probability that the error signal was received as a message word without subsequent knowledge of the sender. The second term represents the probability that knowing the mistake, the sender transmits an erasure but has somehow confirmed the mistake instead. The third term represents the probability that having erased the mistake, the sender has transmitted a second erasure which somehow fails to erase the word which he intended to erase to begin with. Equation 35 can be written as:

$$P_d(c) = \frac{\sum_{ijk}{}' P_{x_\emptyset}(y_{ijk})\, P_{Y_i}(Z_\emptyset)\, P_c(c)}{1 - [1+P_{x_\emptyset}(y\emptyset)\sum_{ijk}{}' P_{x_\emptyset}(y_{ijk})\, P_{Y_i}(Z_\emptyset)} \qquad (36)$$

The above results can be summarized in the form of a theorem:

Theorem II In a iterative discarding system, the error probabilities of the confirmation - denial signal are successively reduced by the iterative process. If the feedback channel is error free, and if the feedback group containing the erasure signal contains it as the one and only signal, the error probabilities decay to zero at approximately an exponential rate with each subsequent confirmed signal. Otherwise they converge rapidly to constants. The probability of failure to finally erase an erroneous word is approximately

$$P_d(c) = \frac{\left\{\sum'_{ijk} P_{x_\phi}(y_{ijk})\, P_{Y_i}(Z_\phi)\right\} \times P_c(c)}{1-[+P_{x_\phi}(y_\phi)] \sum'_{ijk} P_{x_\phi}(y_{ijk}) P_{Y_i}(Z_\phi)}$$

Equivocation of a Confirmed Word

In a iterative discarding system, a finally accepted word has the following background:

1. It has been confirmed.

2. In its previous history, it could have been wrongly transmitted and erased. However, such occurance does not change its equivocation in any way. The rejected word is totally discarded and the receiver is not capable of realizing whatever information left in it.

In the derivation of Theorem 1, equation 11 is perfectly general and will express the equivocation of a finally accepted message word if $P_{ii'}$ is interpreted as the iterated conditional probability that the feedback word $Y_{i'}$ will eventually be confirmed, when $x_{ij} \in X_i$ was originally transmitted.

Theorem III The equivocation of an accepted message word is

$$H_{cy_{i'j'k'}}(x) = \log P_{y_{i'j'k'}}(c) - \sum_i \sum_j P_{cy_{i'j'k'}}(x_{ij}) \left\{\log P_{ii'} + \log P_{y_{i'j'k'}}(x_{ij})\right\}$$

where $P_{ii'}$ is the iterated conditional probability that the feedback word $Y_{i'}$ will eventually be confirmed, when $x_{ij} \in \{X_i\}$ was originally transmitted. $P_{y_{i'j'k'}}(c)$ and $P_{cy_{i'j'k'}}(x_{ij})$ are respectively

$$P_{y_{i'j'k'}}(c) = \sum_i \sum_j P_{ii'} P_{y_{i'j'k'}}(x_{ij})$$

$$P_{cy_{i'j'k'}}(x_{ij}) = \frac{P_{ii'} P_{y_{i'j'k'}}(x_{ij})}{P_{y_{i'j'k'}}(c)}$$

Corollary I If the feedback channel is error free, and if the feedback group containing the erasure signal contains it as the one and only signal, the equivocation of an accepted message word is

$$H_{cy_{i'j'k'}}(x) = \log \sum_j P_{y_{i'j'k'}}(x_{i'j}) - \sum_j \frac{P_{y_{i'j'k'}}(x_{i'j}) \log P_{y_{i'j'k'}}(x_{i'j})}{\sum_{j''} P_{y_{i'j'k'}}(x_{i'j''})}$$

provided that the message word is followed by a sufficient number of accepted message words.

In other words, the equivocation left in such a system is equal to the uncertainty left among the signals within the same group. The Corollary follows directly from the Theorem since in such a system, the iterated $P_{ii'}$ is equal to $\delta_{ii'}$.

Corollary II When the error free feedback channel has a capacity at least as great as the direct channel, information may be sent error free at a rate equal to or less than the information rate of the direct channel. The equality holds if there is no left over information in the rejected message words.

Corollary II follows from Corollary I of Theorem I and Corollary I of Theorem III.

III Examples of Iterative Discarding Systems

1. Error Free Repetitive Feedback

Let us assume a channel with N symmetrical message positions per digit. The probability that a digit is received incorrectly is p. One of the N positions will be used as an "erasure".

From the preceding section the final message will be error free, but the percentage of informative digits that get through is

$$m = (1 - 2p) \qquad (37)$$

Equation 37 is due to the fact that no matter how an error occurs, two digits are nullified with each error occurance. The information rate

per digit is

$$R_f' = m \log (N-1) - (1-2p) \log (N-1) \qquad (38)$$

The information rate of the direct channel alone is

$$R_o' = \log N - E_o \qquad (39)$$

where E_o is the equivocation per digit and depends upon the error distribution. If equal distribution among the N-1 states is assumed

$$E_o = - p \log \frac{p}{N-1} - (1-p)\log(1-p) \qquad (40)$$

The net loss in information rate due to feedback is

$$\Delta R' = R_o' - R_f' = \log N + p \log \frac{p}{N-1} + (1-p)\log(1-p)$$

$$- (1-2p)\log(N-1) \qquad (41)$$

Equation 35 can be written as

$$\Delta R' = p \log p + (1-p)\log(1-p)$$

$$+p \log N + (1-p)\log \frac{N}{N-1} \qquad (42)$$

Let p_1 be defined as $\frac{1}{N}$, then

$$\Delta R' = p \log p/p_1 + (1-p)\log \frac{1-p}{1-p_1} \qquad (43)$$

To evaluate $\Delta R'$, let us determine its stationary point

$$\frac{d\Delta R'}{dp_1} = - \frac{p}{p_1} + \frac{1-p}{1-p_1} = \frac{p_1-p}{p_1(1-p_1)} \qquad (44)$$

Hence $\Delta R'$ has one and only one stationary point at $p_1 = p$. The stationary point is a minimum since $\frac{d\ R'}{dp_1}$ is negative for $p_1 < p$ but positive for $p_1 > p$. Therefore $R' \geq 0$. The equality sign holds only if $p_1 = p$ or $Np = 1$.

If $Np = 1$ two conditions are met:

(a) The rate provided for confirming information,

$$\log N - (1-p)\log(N-1)$$

is equal to the information that must be confirmed

$$-p \log p - (1-p)\log(1-p).$$

(b) There is no residue information in a reject word. Once a word is rejected, its probability of being correct is the same as that of the erasure being wrong. It is $p = \frac{1}{N}$.

2. Error Free Parity Check Feedback

Both the direct transmission channel and feedback channel are composed of binary digits. The feedback channel is error free but of very low capacity. For every m digits sent through the direct channel, 1 digit is reported to check its parity.

p = probability of error per digit, direct channel

p' = probability of odd errors, which will be detected and erased.

p" = probability of even errors which will be passed unnoticed.

One of the 2^m messages per group will be used as an erasure.

p_a = probability of acceptance = $1-2p'$

p_e = probability of error in accepted message groups $\simeq \frac{p''}{1-2p'}$ (45)

Equation 45 is only approximately true. It is assumed that the error probability p" is small such that the probabilities of two errors making it correct or making only one error in the end, etc. are neglected. For more accurate calculations, equation 36 should be used.

C_f = channel capacity per digit with feedback

$$= (1-2p') \frac{1}{m} \log (2^m - 1) \qquad (46)$$

$$p' = \sum_{\substack{i=1 \\ \text{odd}}}^{i \leq m} p^i (1-p)^{m-i} \frac{m!}{i!\ (m-i)!}$$

$$= \left\{ [1-p) + p]^m - [(1-p) - p]^m \right\} /2$$

$$p' = \frac{1}{2}[1 - (1-2p)^m] \qquad (47)$$

$$p'' = \sum_{\substack{i=2 \\ \text{even}}}^{i \leq m} p^i (1-p)^{m-i} \frac{m!}{i!\ (m-i)!}$$

$$= \left\{ [(1-p) + p]^m + [(1-p) - p]^m \right\} /2$$

$$- (1-p)^m$$

$$p'' = \frac{1}{2}\left[1 - 2(1-p)^m + (1-2p)^m\right] \qquad (48)$$

From equations 45, 46, 47, and 48 p_e and C_f can be calculated.

3. Noisy, Parity Check Feedback

In this case, probabilities of odd and even errors will include the feedback digit. For simplicity, let us assume that the feedback digit has the same error probability p. Equations (45) and (46) will remain unchanged, and equations (47) and (48) become

$$p' = \frac{1}{2}\left[1 - (1-2p)^{m+1}\right] \qquad (49)$$

$$p'' = \frac{1}{2}\left[1 - 2(1-p)^{m+1} + (1-2p)^{m+1}\right] \qquad (50)$$

4. Error Free Hamming Check Digit Feedback

For every m digits sent through the direct channel, c digits will be fed back. Out of these, 1 digit is for parity check on the m digits only and the remaining c-1 digits are for single error correcting double error detecting, or triple error detecting. Since there is no need to check on the c-1 digits themselves,

$$2^{c-2} - 1 < m \leq 2^{c-1} - 1 \qquad (51)$$

The inequality 51 determines the number of required feedback digits.

Using this code up to 3 errors will be detected and erased. If there are four errors, the only cases which are not detected are the ones which have the same check digits, i.e., if m=7, the sequences 0000000 and 0111100 have same check digits,0000. Since with three arbitrarily placed error digits, the position of the fourth digit is fixed by the check digits, (for equality of check digits), a condition which always exists if

$$m = 2^{c-1} - 1$$

but sometimes does not exist if

$$m < 2^{c-1} - 1,$$

the probability of four errors is

$$p_4 < \frac{m(m-1)(m-2)}{3!}p^4 (1-p)^{m-4} \qquad (52)$$

Five errors will always be detected by the parity check. The probability of six errors is much smaller than p_4. Hence the probability of an unnoticed error is

$$p'' \approx \frac{m(m-1)(m-2)}{6}p^4 (1-p)^{m-4} \qquad (53)$$

The probability of being correct is $(1-p)^m$. Therefore the probability of a detected error is

$$p' \approx 1 - (1-p)^m - \frac{m(m-1)(m-2)}{6}p^4(1-p)^{m-4} \qquad (54)$$

p_e and C_f can be calculated from equations 45, and 46, using the values of p' and p'' from equations 53, and 54.

If direct repetition were used, the error probability would be of the order p^2, instead of p^4 as is obtained with feedback.

5. Noisy, Hamming Check Digit Feedback

In this case, the c feedback digits are checked together with the m direct information digits. Hence

$$2^{c-2} - 1 < m + c - 1 \leq 2^{c-1} - 1$$

or

$$2^{c-2} < m + c \leq 2^{c-1} \qquad (55)$$

If the same error probability p is assumed for feedback digits as for the information digits, p'' and p' become respectively:

$$p'' \approx \frac{(m+c)!}{3!\,(m+c-3)!}p^4 (1-p)^{m+c-4} \qquad (56)$$

$$p' \approx 1 - (1-p)^{m+c} - \frac{(m+c)!}{3!\,(m+c-3)!}p^4(1-p)^{m+c-4} \qquad (57)$$

Equations 45, 46, 56, and 57 give values of p_e and C_f.

One case of special interest is that of m=4. Inequality 55 gives c=4. The possible combinations are as follows:

Direct Message Digits				Feedback Digits			
7	6	5	3	4	2	1	Parity
0	0	0	0	0	0	0	0
0	0	0	1	0	1	1	1
0	0	1	0	1	0	1	1
0	0	1	1	1	1	0	0
0	1	0	0	1	1	0	1
0	1	0	1	1	0	1	0
0	1	1	0	0	1	1	0
0	1	1	1	0	0	0	1
1	0	0	0	1	1	1	0
1	0	0	1	1	0	0	1
1	0	1	0	0	1	0	1
1	0	1	1	0	0	1	0
1	1	0	0	0	0	1	1
1	1	0	1	0	1	0	0
1	1	1	0	1	0	0	0
1	1	1	1	1	1	1	1

From the above table it is seen that

1. There is a one to one correspondence between the direct message and the feedback message. In terms of the general description of information feedback, each feedback group contains only one message.

2. The distance d(x,y) between two direct messages x and y and the distance d (x', y') between the two corresponding feedback messages x' and y' are related as follows:

d(x,y)	d(x',y')
1	3
2	2
3	1
4	4

The general philosophy of noisy feedback appears to be:

"Select the corresponding feedback messages such that if two messages are close together in signal space, their corresponding feedback messages are far away in the signal space, and vice versa".

Conclusion

In the above, the process of information feedback is formulated in a relatively general form, and basic relations governing information rate and reliability are derived.

Typical examples are given which show that the feedback process is by far more effective in improving reliability compared to direct coding. When a feedback channel of the same or less error probability is available, equivocation can be effectively reduced without laborous coding. The loss in information rate is small. In the limiting case of error free feedback with sufficient capacity for simple repetition, error free transmission can be obtained at little or no cost of information rate. When the feedback channel is not reliable, it can be coded such that substantially less but comparably reliable information will be reported to the sender and much of the advantages of feedback can still be realized.

However more questions are raised than answered. Just to name a few: Could predictive feedback be used to reduce redundancy and to increase communication rate? What are the criteria of optimum coding of the feedback channel in case it is noisy? At what signal to noise ratio could direct transmission channel be uncoded without signigicant loss of information rate? Is there any advantage in a compound feedback process in which both decision feedback and information feedback are used? It is hoped that this paper will focus attention on the problems related to information feedback so that they will be answered by future investigations.

Acknowledgment

This work has been sponsored by the Air Force Cambridge Research Center, Air Research and Development Command, Cambridge, Mass., under contract number AF19(604)1049.

The writer is indebted to his colleagues Mr. Frank J. Bloom and Mr. Bernard Harris. The study of feedback systems was undertaken at the suggestion of Mr. Bloom who also directs this research project. Mr. Harris has done an excellent job in editing portions of the quoted reports[2] from which substantial sections of this paper are taken directly.

References

1. The phrase "in part" is in the Shannon sense of choice or uncertainty, that the feedback message reduces but does not eliminate the uncertainty of the message which has been received. See C.E.Shannon and W.Weaver"The Mathematical Theory of Communication" pp.18-22, University of Illinois Press, 1949.

2. "Supplementary Notes on Evaluation Theory For Communication Systems", Seventh Quarterly Report, September 15, 1955 to December 15,1955; and First Scientific Report, January 15, 1956 to March 15, 1956. Submitted to Air Force Cambridge Research Center by Research Division, College of Engineering, New York University.

3. R. W. Hamming, "Error Detecting and Error Correcting Codes," Bell System Technical Journal, 29, pp.147-160 (1950)

4. For example of iteration in a direct coding scheme, see Peter Elias "Error-Free Coding" Trans. IRE, Information Theory 4 (1954) pp. 29-37.

A LINEAR CODING FOR TRANSMITTING A SET OF CORRELATED SIGNALS

by

H. P. Kramer and M. V. Mathews
Bell Telephone Laboratories, Incorporated
Murray Hill, N. J.

ABSTRACT A coding scheme is described for the transmission of n continuous correlated signals over m channels, m being equal to or less than n. Each of the m signals is a linear combination of the n original signals. The coefficients of this linear transformation, which constitute an m x n matrix, are constants of the coding scheme. For the purpose of decoding, the m signals are once more combined linearly into n output signals which approximate the input signals. The coefficients of the coding matrix which minimize the sum of the mean square differences between the original signals and the reconstructed ones are shown to be the components of the eigenvectors of the matrix of the correlation coefficients of the original signals. The decoding matrix is the transpose of the coding matrix.

As an example, the coding scheme is applied to a channel vocoder in which speech is transmitted by means of a set of signals proportional to the speech energy in the various frequency bands. These signals are strongly correlated, and the coding results in a substantial reduction in the number of signals necessary to transmit highly articulate speech.

The coding theory can be extended to include the minimization of the expectation of any positive definite quadratic function of the differences between the original and reconstructed signals. In addition, if the signals are Gaussian, the sum of the channel capacities necessary to transmit the transformed signals is shown to be equal to or less than that necessary to transmit the original signals.

INTRODUCTION

The coding scheme described in this paper can be used for the transmission of any set of correlated signals. As an example throughout the paper, the transmission of the energy signals in the Channel Vocoder will be considered. However, the scheme is in no way limited to vocoders.

The vocoder is a device developed by H. W. Dudley[1]* for reducing the bandwidth necessary to transmit speech signals. In the vocoder the speech is passed through contiguous band pass filters and the outputs of these filters are rectified and passed through low pass filters. The resulting transmission signals are measures of the energy of the original speech in the various frequency bands. These energy signals can then be transmitted over separate channels and the speech reconstructed at the receiving terminal by a modulation process.

Figure 1 is a sample of a typical set of vocoder energy signals for a 10 channel vocoder. As can be seen from the figure, the various energy signals are highly correlated and consequently some coding should exist which removes this correlation and results in an even further reduction of channel capacity or alternatively more efficient exploitation of the vocoder scheme. One such coding was suggested by the following line of reasoning: Since the signals are so well correlated a desirable procedure might be to send the first signal and the difference between the first signal and the second signal, the first signal and the third signal, etc., these differences being small. Perhaps if the correlation between two channels were sufficiently high, the difference signals for these two channels would be small enough so it could be neglected and thus the total number of channels necessary to send would be reduced by one.

The procedure of sending difference signals suggests transmitting linear combinations of the signals:

$$x_i = \sum_{j=1}^{n} A_{ij}\, e_j \qquad i = 1 \ldots m \qquad (1)$$

In equation 1 the e_j signals are the original time varying vocoder signals, the x_i signals are the time varying transmission signals and the A_{ij} coefficients are constants of the coding. At the receiving terminal the e_j signals are constructed by a second linear transformation

$$e'_i = \sum_{j=1}^{m} B_{ij}\, x_j \qquad i = 1 \ldots n \qquad (2)$$

---

* Superscript numbers refer to references at the end of the paper.

in which e'_i is the reconstructed e_i signal. A diagram of an entire transmission system is shown in Fig. 2. If the number m of x_i signals equals the number n of e_i signals, then the e_i signals can be reconstructed exactly at the receiving terminal by making the matrix of B_{ij} coefficients inverse to the matrix of A_{ij} coefficients. If some signals have been neglected so that $m < n$ then the e_i signals can be reconstructed only approximately at the receiving terminal and the question arises what choice of A_{ij}, B_{ij} coefficients will minimize some measure of the error so introduced. The error measure considered here is the sum of the mean square errors given by the relation.

$$M = \sum_{i=1}^{n} \overline{(e_i - e'_i)^2} \tag{3}$$

As might be expected this selection of error criterion yields a simple solution to the minimization problem, but in addition it has been justified experimentally to some extent in the case of the vocoder.

The exact formula for the choice of coefficients will be discussed in the next section. However, the heuristic basis for this choice can be pointed out here. At each instant in time the values of the n channel signals can be considered as the coordinates of a point in n dimensional space. For example, a two channel vocoder can be represented by the two dimensional space of Figure 3. As time progresses the point will trace out a pattern as shown in Figure 3. Because the signals are well correlated this pattern will tend to concentrate in some particular area. The linear transformation specified by equation 1 amounts to a rotation of coordinate axes on Figure 3 to a new set of axes x_1, x_2. If the rotation is chosen so the x_1 axis corresponds to the major dimension of the pattern then the x_1 signal will contain most of the information about e_1 and e_2. Consequently, neglecting the x_2 signal will result in a minimum error. This procedure will be stated exactly for the n dimensional case in the next section.

OPTIMUM TRANSFORMATION COEFFICIENTS

The mean square error given by equation 3 in the previous section is a function of the A_{ij}, B_{ij} coefficients and of the correlation coefficients R_{ij} of the channel signals where

$$R_{ij} = \overline{e_i e_j} \ . \tag{4}$$

In the Appendix it is shown that to minimize the mean square error when transmitting m ($< n$) signals A_{ij} should be the jth component of the ith normalized eigenvector of the matrix (R_{ij}). These conditions may be specified by the equations

$$L_i \ A_{ij} = \sum_{k=1}^{n} R_{jk} \ A_{ik} \tag{5}$$

and

$$\sum_{j=1}^{n} A_{ij} \ A_{kj} = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} \tag{6}$$

where L_i is the eigenvalue associated with the eigenvector and the eigenvalues are arranged in order of decreasing magnitude.

$$L_1 \geq L_2 \geq \cdots \geq L_n \tag{7}$$

The B_{ij} coefficients are related to the A_{ij} coefficients by

$$B_{ij} = A_{ji} \ . \tag{8}$$

The error is the sum of the eigenvalues for the n - m channels which have been neglected,

$$M = \sum_{i=m+1}^{n} L_i \ . \tag{9}$$

The optimization procedure can easily be extended to include the minimization of the average of any positive, definite quadratic form of the individual channel errors.

CHANNEL CAPACITY REDUCTION

In addition to the error minimization properties described in the previous section, the eigenvector transformation can be shown to yield a saving in channel capacity if the e_i signals are Gaussian. The computation of channel capacity is complicated by the fact that e_i and x_i are continuous signals, and their source rates are infinite unless a fidelity criterion is used. However, if the mean square fidelity criterion M (see Eq. 3) is applied to the total transmission system, the minimum sum of the e_i source rates which satisfies the M criterion can be shown to be equal to or greater than the minimum sum of the x_i source rates which satisfies the M criterion. This result may be interpreted as meaning that if the e_i signals are transmitted over n independent channels, the sum of the channel capacities necessary to meet the M criterion is equal to or greater

than the corresponding sum for the x_i signals.

The existence of a channel capacity saving can be established even when no channels are neglected ($m = n$). The saving for $m = n$ is most simply shown when the x_i signals have flat band-limited spectra, in which case the source rate of an x_i signal may be written[2]

$$R_{x_i} = W \log \frac{\overline{x_i^2}}{\overline{n_i^2}} \qquad x_i^2 \geq \overline{n_i^2} \tag{10}$$

where W is the bandwidth and $\overline{n_i^2}$ is the allowable mean-square transmission error. The sum of the x_i source rates is

$$R_x = \sum_{i=1}^{n} W \log \frac{\overline{x_i^2}}{\overline{n_i^2}} \tag{11}$$

By summing the effects on e_i of the n_i errors the resulting M fidelity criterion can be shown to be

$$M = \sum_{i=1}^{n} \overline{n_i^2} \tag{12}$$

If the $\overline{n_i^2}$ errors are distributed among the various channels so as to minimize R_x while satisfying Eq. (12) the resulting distribution has equal errors in each channel. The minimum value of R_x so obtained is

$$R_x = \sum_{i=1}^{n} W \log \frac{n\,\overline{x_i^2}}{M} \tag{13}$$

Similarly, the minimum sum of the e_i source rates which satisfy Eq. (12) is

$$R_e = \sum_{i=1}^{n} W \log \frac{n\,\overline{e_i^2}}{M} \tag{14}$$

Because the x_i signals are uncorrelated and (A_{ij}) is an orthonormal matrix, $\overline{e_i^2}$ can be written in terms of $\overline{x_i^2}$ as

$$\overline{e_i^2} = \sum_{j=1}^{n} A_{ji}^2 \, \overline{x_j^2} \,. \tag{15}$$

If Eq. (15) is substituted into Eq. (14), the difference in channel capacities, $R_x - R_e$, may be written

$$R_x - R_e = \sum_{i=1}^{n} W \log \frac{\overline{x_i^2}}{\sum_{j=1}^{n} A_{ji}^2 \, \overline{x_j^2}} \tag{16}$$

That difference is less than or equal to zero for any orthonormal (A_{ij}) matrix is a direct result of the convexity of the logarithm function[3].

A similar proof may be carried out for the more general case where the e_i signals are Gaussian but with any arbitrary spectra.

Under some circumstances the allowable error for a signal, $\overline{n_i^2} = M/n$, may be greater than the signal $\overline{x_i^2}$. The required channel has essentially zero capacity and may be neglected, thus reducing the number of channels to be transmitted.

EXPERIMENTAL TEST OF CODING SCHEME

The coding procedure which has been described was evaluated by applying it to a 16-channel vocoder. The correlation coefficients R_{ij} were obtained by time averaging $e_i e_j$ over a four minute sample of speech to the vocoder from eight different male speakers. The eigenvectors associated with the R_{ij} matrix and the associated eigenvalues were computed on an IBM 650 computer using an iterative routine[4]. The eigenvalues so obtained were as follows: 9.57, 1.36, 0.72, 0.55, 0.35, 0.34, 0.31, 0.23, 0.20, 0.10, 0.07, 0.05, 0.01, 0, 0, 0. The theory shows that the error committed by omitting a channel is equal to the eigenvalue associated with that channel. As a consequence, in order to obtain a substantial reduction in the number of channels, the eigenvalues should have a large spread. This condition exists for the vocoder data.

Systems with one, three, six and ten transmission channels were realized and their performance is roughly as expected. The one-channel system was articulate only on very familiar text such as "Mary had a little lamb ...". The three-channel system was estimated to have a sentence

articulation rate better than 50% for unfamiliar text. The six-channel system was almost completely understandable though its quality was substantially less than that of the 16-channel vocoder. The ten-channel system was of good quality.

The quality of the 10-channel system was judged to be better than that of an existing 10-channel vocoder with which direct comparisons could be made. This result leads to the conclusion that to achieve better quality with a given number of channels, a system should be used consisting of a many channel vocoder plus a matrix transformation to attain the required number of transmission channels.

Equations (14) and (15) were applied to evaluate the upper limits of the source rates of the e_i and x_i signals by assuming both sets of signals are Gaussian with flat spectra, band limited at 15 cps., and that the allowable mean square deviation $\Sigma \overline{e_i^2}/M$ is 100:1. These figures have been established as being approximately correct for existing vocoders. A capacity of 1500 bits/sec. will transmit the e_i signals directly and a capacity of 890 bits/sec. will transmit the x_i signals. Thus a saving of a little more than 1/3 of the original channel capacity is achieved by the matrix.

Certain discrepancies between the observed performance of the transmission system and the expected performance based on the eigenvalues can be attributed to the difference between the mean square fidelity criterion and that used by the ear. For example, preliminary articulation measurements indicate a relatively lower articulation score for certain short, but important sounds, such as stops (p, b, t, etc.). Because these sounds are of such short duration, their spectra would not be well represented in the time average used to evaluate $\overline{e_i e_j}$, thus large errors could be expected in their transmission.

CONCLUSIONS

A linear coding scheme has been developed for the transmission of n correlated signals over m channels. The coding is an optimum way of reducing the number of transmission channels where fidelity is measured by the sum of the mean square errors. In addition the coding results in a reduction of the sum of the channel capacities necessary to send the signals over independent channels.

The increased efficiency of the coding procedure may be utilized in two ways in a vocoder. Either speech of substantially the same quality can be transmitted over fewer channels or higher quality speech can be sent over the same number of channels.

The coder involves no memory and thus can be instrumented very simply with a resistance network plus sign inverting amplifiers. This simplicity is a practical advantage over a more efficient coding scheme which requires more complicated instrumentation.

The minimization is of theoretical interest since the mean square error is not a linear function. Thus the minimization has been carried one step further than in most mean square procedures, which essentially minimize only linear functions. In signal theory terminology this result is equivalent to saying that an optimum way of both decomposing and recomposing the original signals has been developed instead of the usual optimization which includes only an optimum recomposition after an arbitrary decomposition.

Appendix

The problem at hand is to minimize the average error,

$$M = \overline{\sum_{k=1}^{n} \left(e_k - \sum_{i=1}^{m} \sum_{j=1}^{n} B_{ki} A_{ij} e_j \right)^2} \qquad (17)$$

by appropriately choosing an m x n matrix A and an n x m matrix B. We begin by determining matrix coefficients A_{rs} and B_{rs} that will result in m a stationary value for M.

$$\frac{\partial M}{\partial B_{rs}} = 2 \left(e_r - \sum_{i=1}^{m} \sum_{j=1}^{n} B_{ri} A_{ij} e_j \right) \left(\sum_{t=1}^{n} A_{st} e_t \right) \qquad (18)$$

$$= 2 (AR)_{sr} - 2 (ARA^T B^T)_{sr}$$

$$\frac{\partial M}{\partial A_{rs}} = 2 \sum_{k=1}^{n} \left(e_k - \sum_{i=1}^{m} \sum_{j=1}^{n} B_{ki} A_{ij} e_j \right) (B_{kr} e_s) \qquad (19)$$

$$= 2 (RB)_{sr} - 2 (RA^T B^T B)_{sr}$$

Here, R is the matrix of correlation coefficients, $R_{ij} = \overline{e_i e_j}$, which is symmetric and positive and may be considered positive definite without loss of generality. The necessary conditions for stationarity stated in matrix notation, are from Eqs. (18) and (19).

$$AR = ARA^T B^T \qquad (20)$$

$$RB = RA^T B^T B \qquad (21)$$

With the substitution BA = C, Eqs. (20) and (21) can be put in the form.

$$CR = CRC^T \tag{22}$$

$$RC = RC^TC \tag{23}$$

By transposing both members of Eq. (22) we find that

$$(CR)^T = (CRC^T)^T = CRC^T = CR$$

and thus

$$RC^T = CR \tag{24}$$

Multiplying Eq. (23) through by R^{-1} and taking transposes, we find that

$$C = C^TC = C^T \tag{25}$$

Combining (25) with (24) yields

$$RC = CR \tag{26}$$

and Eq. (25) can be written

$$C = C^2 \tag{27}$$

Recalling that the trace of a matrix is the sum of its diagonal elements, we can write Eq. (17) in the form

$$M = Tr[\,(I - C)(I - C^T)R\,]$$

Since R is symmetric, there exists a non-singular matrix U with $U^{-1} = U^T$ such that

$$U^TRU = L \tag{28}$$

is diagonal. Eq. (26) implies in addition that

$$U^TCU = P \tag{29}$$

is also diagonal. Let the rank of C be $k \leq m$. Then the rank of P is also $k \leq m$ and

$$P^2 = U^TCUU^TCU = U^TC^2U = U^TCU = P \tag{30}$$

implies that the diagonal elements of P are either zero or one and therefore there are $k \leq m$ one's on the diagonal of P.

Now

$$M = Tr[\,(I - C)R\,] = Tr[\,U^T(I - C)UU^TRU\,] = \Sigma' L_i \tag{31}$$

where the summation extends over n - k of the eigenvalues of R. M is therefore minimized by letting k = m and choosing P so that its non-vanishing terms correspond to the m largest eigenvalues of R.

The matrix C that achieves this minimum is given by

$$C = UKU^{-1}. \tag{32}$$

A check will easily show that if $[U]_k$ denotes the k^{th} column matrix of U,

$$R[U]_k = L[U]_k, \tag{33}$$

so that $[U]_k$ is an eigenvector of R corresponding to the eigenvalue of L_k. The property $U^{-1} = U^T$ implies that $[U]_k$ is normalized. And thus, finally, the choices

$$A = KU^T \tag{34}$$

$$B = UK \tag{35}$$

while not yielding unique A and B matrices, achieve the desired minimum value of M.

The above proof was first given by the author under restrictive condition on A and B. D. Slepian was able to remove these and the present proof owes its generality to him.

References

1. H. Dudley, Remaking Speech, J.A.S.A., May 1939.
2. C. E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, Vol. 27, July, October, 1948.
3. G. Polya and G. Azego, Aufgaben and Lehrsatze aus der Analysis, Dover, 1945 N.Y.C.
4. R. T. Gregory, Computing Eigenvectors and Eigenvalues of a Symmetric Matrix on the ILLIAC, Mathematical Tables and Other Aids to Computation, Vol. 7, 1953, pages 215-220.

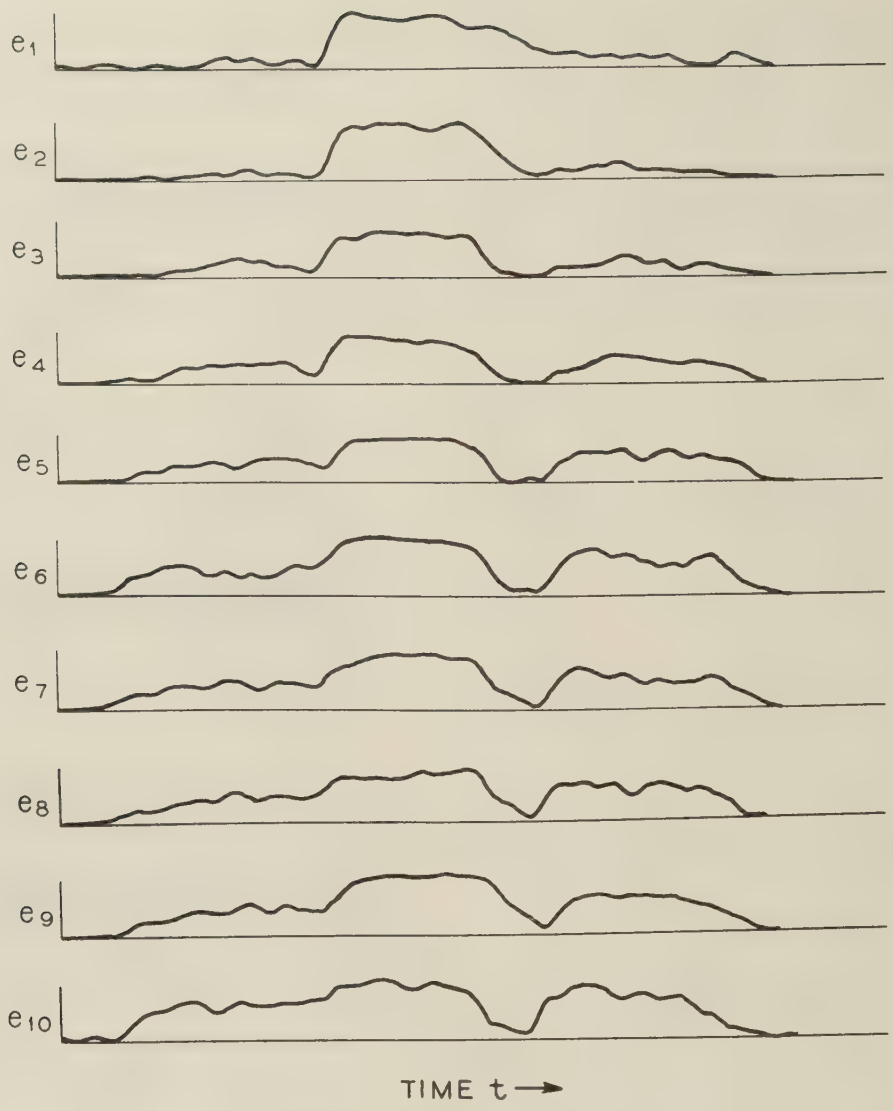


Fig. 1 - Energy signals in ten channel vocoder.

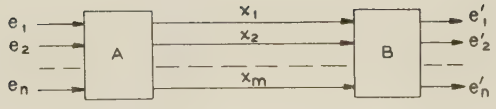


Fig. 2 - Complete transmission system.

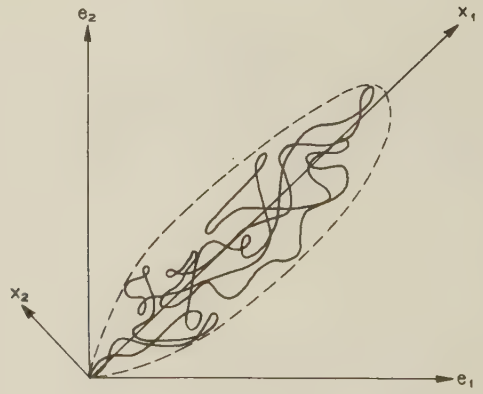


Fig. 3 - Representation of channel vocoder.

ON an APPLICATION of SEMI GROUPS METHODS TO SOME PROBLEMS in CODING

By M.P. Schützenberger
(C.N.R.S. Paris)

0. Introduction.

The current paper deals with a chapter in what could be called communication theory in extensive form : it starts with extremely restricted structures and it stops where begins the canonical problem of optimalisation. It even ends sooner for no full use of the definitions is made and the main ergodic theorem is stated without proof.

Actually the nature itself of the question under study has commanded these restrictions together with the architecture of the paper : we give a abstract model of some sort of language and we try to show how semi group concepts apply fruitfully to it with the hope that some of them may be at least of stimulating interest to specialists working on natural languages.

As frequent in the field of cybernetic, the mathematics involved even if quite simple are far away from classical analysis and, indeed, many of the necessary tools had to be sharpened especially for the purpose.

Thus the paper is twofold : in a first part the model and its main properties are discussed at a concrete level on the simplest cases : the coding and decoding with length bounded codes. In a second part a selection of theorems are proved whenever the necessary semi group theoretic preliminaries are not exacting. The link along this tail of appendices is the theory developped verbally in the first part. Finally a special chapter provides a bridge toward probalistic applications.

It is proper at this place to acknowledge the contributions of three authors who influenced deeply the building of the theory :

Sardinas and Patterson[1] who discussed first on a logical basis the general coding process.

B. Mandelbrot[2] who recognised and studied extensively the role of "word units" in communication theory and related the problem to Feller's recurrent events.

P. Dubreuil[3] and his school whose pionnering work on discrete semi groups has provided many basic concepts and arguments as it will be seen below.

Part I

1. Preliminary definition of a discrete semi group language :

We shall be concerned with the two basic sets of communication theory :

The set of all messages which may possibly be sent.

The set of all signals available for transmission along the line.

The main feature of the theory is the postulationnal requirement that the signals as well as the messages pertain both to some common class of structures so that coding and decoding not only be inverse operations but far more generally, be special instances of a quite broad new process, that of translation.

This identity of structure itself between two sets is a result from the basic restriction that they develop homogeneously in time - or more accurately that both admit a common partial order and composition operation.

That such requirements are rather stringent is clearly seen by the exemple of photography (two exposures give rarely a result which is, in any sense, equivalent to a third one) or even by harmonic modulation where Fourier transform exchanges so well time and frequency that finite signals cannot be fully adequate.

On the other hand, languages either spoken, written or gesticulated are somewhat akin with our consideration, and we shall use the name of "discrete semi group languages" (d.s.g.l.) for naming the elemental concepts of our study.

The definitions below are quite general and as said before, no full use of them will be made here - very little gain in simplicity would be achieved by using more restrictive ones.

DEFINITIONS :

I. A discrete semi group language will be a set Λ of object called "messages" satisfying the following conditions :

I.1. If λ_i and λ_j pertain to Λ so does their "product" $\lambda_k = \lambda_i \lambda_j$ made up of "λ_i" followed by "λ_j" (λ_i will be said a left divisor and λ_j a right divisor of λ_k).

I.2. If λ_l , λ_m and λ_k pertain to Λ and if $\lambda_l = \lambda_i \lambda_j$ and $\lambda_m = \lambda_j \lambda_k$ then $\lambda_i \lambda_m$ is identical with $\lambda_l \lambda_k$.

I.3. The "vacuous message" φ pertains to Λ and satisfies $\varphi \lambda_i = \lambda_i \varphi = \lambda_i$ for all $\lambda_i \in \Lambda$

I.4. There is a sub set Λ_c from Λ called "dictionary" or "basis" whose elements are called "words". Λ_c is such as :

I.4.1. φ does not pertain to Λ_c

I.4.2 for all $\lambda_i \in \Lambda - \varphi$

either $\lambda_i \in \Lambda_c$

either these exist a unique finite set of words $\lambda_{i_1}, \lambda_{i_2}, \ldots \lambda_{i_m} \in \Lambda_0$ with

$$\lambda_i = \lambda_{i_1} \lambda_{i_2} \ldots \lambda_{i_m} .$$

II. Given two d.s.g.l. Λ and M a correspondence θ between the elements of two subsets $\Lambda' \subset \Lambda$ and $M' \subset M$ will be said a translation if it satisfies :

II.1. The correspondence is one to one where ever it is defined.

II.2. If $\lambda_i, \lambda_j \in \Lambda', \theta\lambda_i = \mu_i$, $\theta\lambda_j = \mu_j$, then $\lambda_i \lambda_j \in \Lambda'$ and $\theta(\lambda_i \lambda_j) = \mu_i \mu_j$

II.3. The translation will be said :

Total from Λ to M, if $\Lambda' = \Lambda$.
Subtotal from Λ to M, if for all $\lambda_i \in \Lambda$ there is at least a $\lambda_j \in \Lambda$ such as $\lambda_i \lambda_j \in \Lambda'$.

III. A neat coding of Λ into M will be a translation total from Λ to M and subtotal from M to Λ.

In algebraic form we could reduce our axiomatic to :

I' : Λ is the free discrete semi group generated by Λ_0

II' : A translation is an isomorphism between the sub semi groups $\Lambda' \subset \Lambda$ and $M' \subset M$

III' : A translation is a neat coding if $\Lambda' = \Lambda$ and M' is a subsemigroup of M neat on the right. (Note that "subsemigroup" entails I.1, I.2 and I.3 ; "free" corresponds to unique in I.4.2 , "discrete" to finite at the same place).

2. Practical significance of the axiomatic :

Let us take a simple example in coding :

$\Lambda_0 = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, $M_0 = \{+, -\}$

(M_0 is the usual binary alphabet; Λ is the set of all strings of a finite number of the "elementary messages" λ_i $(i = 1,2,3,4)$ and M is built in the same way with the "letters" + and - .

When coding, we want to establish a correspondence between Λ and some subset M' of M satisfying two conditions :

1) to every $\lambda \in \Lambda$ corresponds at least one $\mu \in M$ ("total" character of the coding)

2) to any distinct $\lambda, \lambda' \in \Lambda$ must correspond distinct $\mu, \mu' \in M'$ in order that the deciphering be free from ambiguity.

A priori any one to one correspondance between Λ and a subset M' from M would do - but usually this could imply that we cannot proceed to the sending of the message before we know it in its totality. So a further practical condition - which is not too easy to formulate rigourously - could be :

3) For a reasonably large number of messages λ the coding is such that for any right multiple λ' of λ (i.e. any $\lambda' = \lambda\lambda''$) the signals μ and μ' have a reasonably long common left divisor μ_1 (i.e. are of the form : $\mu = \mu_1 \mu_2$ and $\mu' = \mu_1 \mu_2'$).

The simplest way of fulfilling these desiderata is to assign to each $\lambda_i \in \Lambda_0$ a string of binary letters μ_i (which very conveniently we may too call a word) and for any sequence $\lambda_{i_1} \lambda_{i_2} \ldots \lambda_{i_m}$ to send the corresponding sequence: $\mu_{i_1} \mu_{i_2} \ldots \mu_{i_m}$.

For example, with the correspondance : $\mathcal{C}_1$:
$\lambda_1 \to + = \mu_1$; $\lambda_2 \to + - = \mu_2$; $\lambda_3 \to - + - = \mu_3$;
$\lambda_4 \to - - = \mu_4$

we would have :

$\lambda_2 \lambda_3 \lambda_1 \lambda_1 \lambda_2 \to + + - - + - + + + - -$.

It is not obvious however how the set M_0' of the words μ_i has to be selected so that decoding be free from ambiguity :

At my knowledge, the question has been raised first and practically solved by Sardinas and Patterson in a pioneering paper(1).

With the help of semi group concepts we may however obtain a deeper insight into their whole procedure which was purely logical :

We are looking for a total translation from Λ to M and it is quite axiomatic that the decoding is unambiguous if and only if the sub semi group M' generated by M_0' is isomorphoric to the free semigroup Λ - or - for short - that M' is a free subsemigroup of M .

Algebraic consequences of this simple remark are to be found in appendix 1.

Now would come a fourth requirement : (admissibility)

4) The length of the words μ_i must be as small as possible in respect of some a priori probability distribution on Λ .

As a matter of fact (4) will be met incidentally, so to say, in wiew of another condition we put in definition III :

That the translation from M back to Λ be sub total :

What this means exactly is that any sequence μ of binary digit be a left divisor of at least one message $\mu' \in M'$ which can be completely and exactly retranslated into Λ .

This condition together with the possibility of one-to-one deciphering implies automatically that the code be unitary (as defined below)(see appendix 0), and admissible in that sense that it meets the optimality requirement (4) in respect of at least one a priori probability distribution of the words. (*)

3. Discussion of the decoding methods : scansion

This being settled we have to look more closely at the decoding.

For avoiding repetition let us observe that Λ does not play any role by itself since the $\lambda_i \in \Lambda_0$ are in a one-to-one correspondance with the words $\mu_i \in M'_0$. So we may perfectly well dispense from mentioning it altogether.

But in order to stress when a given string μ of binary symbols is really a set made up of a sequence of words and not any odd sequence of + and - we shall say that μ is a complete message (for instance : "$+\,+\,-\,-\,+\,-$" $= \mu_2 \mu_3$ is a complete message, but "$+\,+\,-\,-$" is not) and indicate it by enclosing it into two / signs, which shall denote too, end and beginning of the words.

Let us try to decode the following complete message in code $\mathcal{C}_1$:

$/+\,+\,-\,-\,+\,-\,-\,+\,-\,-\,+/$

The only way open is trial and error : the first + may be:

- either μ_1 itself
- either the first letter from $\mu_2 = /+\,+\,-/$

so that we have the choice between :

$/+/+\,-\,-\,+\,-\,-\,+/$ and $/+\,+\,-/-\,+\,-\,-\,+\,-\,-\,+/$

In the first case no further doubt comes in and we are lead to :

$/+/+/-/+/-/+/-/+/ = \mu_1 \mu_1 \mu_4 \mu_1 \mu_4 \mu_1 \mu_4 \mu_1$

In the second we obtain :

$/+\,+\,-/+\,+\,-/+\,-/+ = \mu_2 \mu_1 \mu_2 \mu_3 - +$

Since here $-+$ is left at loose end (strictly speaking) the first translation was the good one, being known that the transmission is over. Observe that if, on the contrary, the signal was the same as before except for an added terminal - digit, the conclusion would be exactly opposite :

$/+\,+\,-/+\,-/+\,-/+\,-/$

is the only fitting "scansion" as we could say by borrowing from prosody this term for its classical flavour.

So the inverse translation from M back to Λ does not look like satisfying very reasonably the above condition 3.

An obvious remedy to it would be to limit still more the set M'_0 . B. Mandelbrot, who has first discussed these problems has distinguished several possibilities :

1) Uniform codes : in which every word has the same length (i.e. number of letters), this criterium giving a direct scansion (examples: all the noise reducing codes introduced so far except for a proposal of "sequential coding" by Peter Elias(4) and some examples by Lemnael(5).)
2) More generally : what we shall call :

   Unitary codes : i.e. codes in which no word is a left divisor of another word (examples : Fano's, Huffman's, Shannon's codes)
3) Natural codes : (introduced by B. Mandelbrot) in which a special letter points out the end of the word (example : most of the spoken or written languages).

   Further, Mandelbrot has shown that any unitary code is, at least asymptotically, as good from the point of view of economy of length as any other one. It could seem futile then to care for more extensive classes were we not prompted by other circumstances - and especially by the threat of a noise.

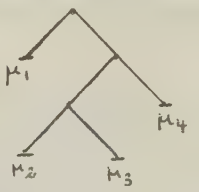
4. Noise absorption and eryodism.

Consider indeed the following code : $\mathcal{C}_2$

$\mu_1 = +\,; \mu_2 = -\,+\,+\,; \mu_3 = -\,+\,-\,; \mu_4 = -\,-$

(which is, parenthetically, just the previous one with the time arrow inverted)

It is unitary all right so that we may represent it by a "tree" in the familiar fashion :



(south west arrow : +
south east arrow : −)

The "neat" condition (subtotality of the translation from M back to Λ) is reflecting itself in the fact that any branch of the tree

(*) If M is the free semi group of all phonemic sequences in English and M' the sub set of all "semantically correct sentences", M' is neat in M .

(For instance :

" /pri wat law cut chur coco feet .."(obtained from King Lear, Act III, scene I, with Tippet's help) is fitted into a complete message in M' by adding : "... and this, Gentlemen, was, may-be, my best example of a semantically void utterance/")

ends with a word (for example the code $\mu_1 = +$; $\mu_2 = -+r$; $\mu_3 = --$ would not be neat since no word nor sequence of words may begin with /- + - ...).

Suppose that we have to decode the sequence :

/- + - - + - + - - + - -

we obtain directly :

/- + - | - + - | + | - - | + | - - |

and we could have written it down extemporaneously without waiting for the end of the transmission.

But if the first digit had been blurred by noise, this straight forward attitude could not be kept : indeed we decipher the uncertain message

/? - - + - + - - + - - ....

either as :

/ + | + | - - | + | - + - / - + - | - ....

either as above :

/ - + - | - + - | + | - - | + | - - / ....

and as long as the message is going on we have no evidence for deciding between this two interpretations. Things nonetheless are not so bad as they look at first glance :

Suppose that the next letters which appear be + + - + - + - ....

so that up to this time the two alternative versions are :

/ + + / - - / + / - + - / - + - / - + + // - + - / + / - ...
/ - + - / - + - / + / - - / + / - - / + / + // - + - / + / - ....

By the seemingly fortuitous fact that in both case the end of a word falls exactly as the same spot (marked // above), the two translations coincidate from this point on and since one of them must be right so is the end of the deciphering - assuming of course that no new error of transmission takes place.

Practically, if such a fact was frequent enough, this would mean that for very low levels of noise, considerable parts of the "meaning" could be preserved. We shall see that this ergodic property (i.e. this relative independence for long sequences of the scansion of the end from that of the beginning) is the rule rather than the exception.

More specifically, for neat codes whose words have all a bounded lenght an apart from three exceptional families there is at least one finite sequence of words - say μ_∞ such that whatever be the initial sequence α, $\alpha\mu_\infty$/ is a complete message. This implies that, when decoding, any blurring or error in α is "absorbed" by μ_∞ and that from the end of μ_∞ on, the scansion starts all right afresh.

Now if the words are given randomly and independently with fixed probabilities, it is clear that the probability for a given sequence not to contain μ_∞ tends with its length exponentially to zero so that any initial error is most likely to have only limited effects.

5. Syntactic equivalence and the fondamental semi groups.

Suppose we be given in code $\mathcal{C}_2$ the following fragment μ from a message :

μ : ... + - - + - - + - + - - ...

By trial and error we see that only three scansions can possibly be fitted to it :

1) ... / + / - - / + / - - / + / - + - / - ...

2) ... + / - - / + / - - / + / - + - / - ...

3) ... + - / - + - / - + - / + / - - / ...

In the same manner the fragment

μ' : ... + - + - + - ...

would give alternatives :

1) ... / + / - + - / + / - ...

2) ... + / - + - / + / - ....

3) ... + - / + / - + - / ...

Disregarding the "meaning" of μ and μ' (i.e. their eventual decoding into the $\mathcal{A}$ language) we may observe that "functionally", so to say, μ and μ' are quite similar :

If the complete message is / $\mu_1 \mu \mu_2$ / , the only possibilities are for each of the three scansions :

1) μ_1 is a complete message and μ_2 starts with -/.. or + + /.. or + - /.. (so as to make use of the /- .. left at the end of μ).

2) μ_1 is ending by ../- + (so as to use .. + /) and μ_2 starts as above.

3) μ_1 ends with ../- (for the sake of... + - /) and μ_2 is a complete message.

Easy check shows that the same applies exactly to μ' and we shall say that μ and μ' are syntactically equivalent (*) ($\mu \equiv \mu'$). Actually both are equivalent to an even simpler fragment :

μ'' = ... + - ...

since this last one admits the same scansions :

1) ../ + / - .. ; 2) ... + / - .... ; 3) ... + - / ...

(*) It is interesting to observe that syntactic equivalence has a direct application to normal linguistics:

If $\mathcal{M}'$ is the set of all sentences grammatically correct :

$\mu_1 \equiv \mu_2$ (approximatively !) if and only if μ_1 and μ_2 pertain to the same grammatical category (for instance in English : both "adjectives" or both "verbs at the third person of the present" etc.)

Now the key point is that for any four finite fragments, μ_1 , μ_2 , μ_3 and μ_4

$\mu_1 \equiv \mu_2$ and $\mu_3 \equiv \mu_4$ implies $\mu_1\mu_3 \equiv \mu_2\mu_4$.

The syntactic equivalence is thus fully compatible with the semi group structure of M and if we consider classes for $\equiv$ (i.e. the subsets of elements from M which are syntactically equivalent between themselves), these classes make a new semi group $\overline{M}$ which is an homomorphic image of M .

$\overline{M} \supset \overline{M}_0$, the fondamental semi group of the coding (f.s.g.) is most usually finite and is easily represented by matrices, but before we explain how, we need still a new concept : that of prefix :

Consider again two fragments μ and μ' but assume, now, that both are beginning at a / mark:

Even if μ and μ' are not syntactically equivalent, it could happen that under this supplementary restriction any further fragment which completes μ into a full message would do the same to μ' :

One could say that " μ and μ' as beginning of messages are syntactically equivalent on the right" (in symbols : $\mu \sim \mu'$)

For example :

$\mu = /--- \cdots$ and $\mu' /+- \cdots$ $/-+-/--/\cdots$ are not in the relation $\equiv$ (since $/-+\mu'' =$ is a complete message although $/-+\mu' = /-++/- \cdots$ is not complete), but $\mu \sim \mu'$ all the same for $\mu\mu''$ is a complete message if and only if

$\mu'' = -/\cdots$ or $++/\cdots$ or $+-/\cdots$

just as well as for μ' .

We call prefixes the classes π_i of fragments for this new relation $\sim$.

For the code $\mathcal{C}_2$, there are three prefixes :

$\pi_1 \ni |\emptyset|$ (words and words only are bringing a $\mu \in \pi_1$ into a complete message.

π_1 contains all the words and its existence is typical of unitary coding).

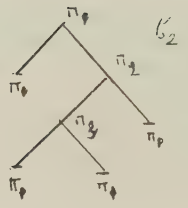
$\pi_2 \ni /- \cdots$ (the corresponding right divisors are $-/\cdots$, $++/\cdots$ and $+-/\cdots$)

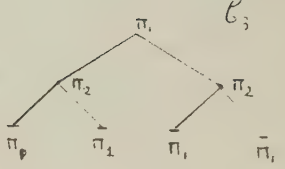
$\pi_3 \ni /-+ \cdots$ (the corresponding right divisor are $+/\cdots$ or $-/\cdots$).

Now if $\mu_1 \sim \mu_2$, one proves that

$\mu_1\mu_3 \sim \mu_2\mu_3$, too, whatever be μ_3

With unitary codes prefixes correspond to nodes of the tree in a one to many fashion : Two nodes being in relation $\sim$ ("pertain to the same prefix") if the subtrees below them are identical. Such things does not occur in our $\mathcal{C}_2$ code (see below), but are quite typical of uniform codes.





In the code $\mathcal{C}_3$ of length 2 ($\mu_1 = ++$; $\mu_2 = +-$; $\mu_3 = -+$; $\mu_4 = --$) there is only two prefixes : one, π_1 , corresponding to complete messages - i.e. to sequences with an even number of letters - and another one, π_2 , corresponding to odd length sequences.

6. Matrix representation of the fundamental semi group.

If we have started reading just at the beginning of the transmission, we may consider at any time t the prefix $\pi(t)$ to which pertain the initial fragment till the t-th letter as a "state" which changes at any new letter received.

For instance - apart from any meaning again - the sequence $/+--+-\cdots$ corresponds to the following sequence of prefixes :

. $\pi_1\ \pi_1\ \pi_2\ \pi_1\ \pi_2\ \pi_3\ \pi_1\ \pi_1\ \pi_c$

It is easy to visualise " + " and " - " respectively as the transition matrices :

	π_1	π_2	π_3
π_1	1	0	0
π_2	0	0	1
π_3	1	0	0

(+)

	π_1	π_2	π_3
π_1	0	1	0
π_2	1	0	0
π_3	1	0	0

(-)

(+ lets π_1 invariant since it is a word. It sends π_2 into π_3 and makes a word from π_3 etc..)

These matrices correspond in a one to one fashion to the elements of the fondament semi group, for instance:

1	0	0
1	0	0
1	0	0

(+ +)

0	0	1
1	0	0
1	0	0

(- +)

(with the usual line by column multiplication) is the matrice given below. What are the matrices corresponding to complete messages ? In the general case they are the matrices of the subsemigroups $\overline{M_d}$ image of M' by the syntactic homomorphism .

But if the code is unitary, $\overline{M'}$ is characterised very nicely, since $\mu \in M'$ implies that μ sends π_1 into itself : $\overline{M'}$ is just the set of the matrices of with 1 in the top left corner.

Further, noise absorption - or ergodic - properties reflect themselves quite directly on this matrix representation.

Suppose that the correct message $/\mu \ldots$ and the perturbated message $/\mu' \ldots$ fall back both at this very time on a common scansion mark $/$. If the prefix corresponding to μ was π and that corresponding to μ' was π', this would mean that the next signals sends both π and π' into $\cdot\pi_1$.

On the matrices this is expressed by the fact that in column π_1 there is two 1's : one in the line π and another one in line π'. In particular μ_∞ is a matrix with 1 everywhere in column π_1.

But this in turn is linked closely with the fact that $\overline{M}$ is a semi group and not a group (whose matrices should all have a single 1 by column).

Consider as a counter example the uniform code with four words :

Its f.s.g. is just the cyclic group of order two, made up of the two elements :

	π_1	π_2
π_1	0	1
π_2	1	0

	π_1	π_2
π_1	1	0
π_2	0	1

(+ or − or any odd length sequence)

(ϕ or ++ or +− or −+ etc. or any even length sequence).

No real absorption takes place for indeed if we had missed the first letter of the transmission and started wrongly scanding from the second letter, the error will obvious go on as long as does the message.

As a matter of fact uniform codes are the only neat codes with a bounded length for words whose f.s.g. is a group. They are the first exceptional non ergodic family.

7. Super coding.

We have given a very general definition of "translation" which suggests the possibility of more complex processes involving not only two but several languages. In the general case, things are a bit confused and we shall restrict ourself to Unitary Neat Coding from K into Λ and from Λ into M.

Suppose for instance that we have the following set up :

K is a d.s.g.l. with words K_i $(1 \leq i \leq 7)$

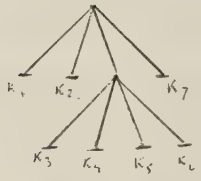
Λ is a d.s.g.l. with words λ_α $(1 \leq \alpha \leq 4)$

M is our familiar binary d.s.g.l.

Each word of Λ is coded in M as in example 2:
$\lambda_1 \to +$; $\lambda_2 \to -++$; $\lambda_3 = -+-$; $\lambda_4 = --$.

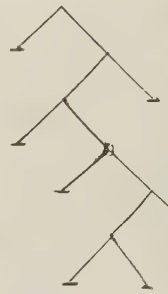
Each word of K is coded by the following sequences $\lambda^{(j)}$ of Λ (for clarity we use upper and lower indices) : $K_1 \to \lambda^1 = \lambda_1$; $K_2 \to \lambda^2 = \lambda_2$; $K_3 \to \lambda^3 = \lambda_3\lambda_1$
$K_4 \to \lambda^4 = \lambda_3\lambda_2$; $K_5 \to \lambda^5 = \lambda_3\lambda_3$; $K_6 \to \lambda^6 = \lambda_3\lambda_4$, $K_7 \to \lambda^7 = \lambda_4$.

This coding is unitary and neat all right and corresponds to the tree :



Now there is again a coding of K into M when every λ^j is written in binary alphabet :
$K_1 \to +$, $K_2 \to -++$, $K_3 \to -+-+$; $K_4 \to -+--++$;
$K_5 \to -+--+-$; $K_6 \to -+----$, $K_7 \to -+---$.

It is not difficult to see that this $K \to M$ coding is unitary and neat. Its tree is given below.



Since we know the importance of fundamental semi groups we would be interested to get at once that ($\overline{\overline{M}}$) of the $K \to M$ process from the other two ($\overline{\Lambda}$ for $K \to \Lambda$ and M for $\Lambda \to M$) or, alternatively, to know the relation between the syntactic equivalences on the bottom structure M : $\equiv(\Lambda$ in respect of $\Lambda \to M$, without K appearing in the picture and $\equiv(K)$ in respect of $K \to M$ with Λ put off from the circuit.

The main result is that :

$\mu_1 \not\equiv \mu_2 \ (\Lambda)$ entails $\mu_1 \not\equiv \mu_2 \ (M)$

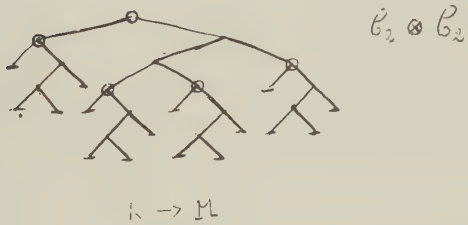
or, if one prefers, that $\overline{M}$ is a homomorphic image of $\overline{\overline{M}}$.

This is rather convenient from a technical point of view for it allows what is called a

filtering. If starting from the assumption that the λ_i are provided independently with fixed probabilities by the source, we discover later on that, actually, they were just building blocks in some higher degree semantic units (sent again independently of each other as a second approximation) we can preserve at least some of the features of our initial approximation.

But the main point for us here lies in another aspect.

Suppose that the $K \to \Lambda$ coding be uniform. in general the $K \to M$ one will not be so, but it will fail to be ergodic just the same, giving us the second of the three exceptional families mentionned above. We shall call such codes "uniformily composed codes". An example is given below :



$\mathcal{C}_2 \otimes \mathcal{C}_2$

$K \to M$

$K \to \Lambda$: uniform of length two

$\Lambda \to M$: our usual $\mathcal{C}_2$

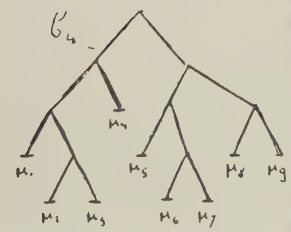
The nodes indicated with a o are the ones corresponding to nodes in the $K \to \Lambda$ coding.

8. Anagrammatic codes.

Let us come now into the last family. For this we produce the following horrible example : $\mathcal{C}_4$

$\mu_1 = +++$; $\mu_2 = ++-+$; $\mu_3 = +++--$; $\mu_4 = ++-$.

$\mu_5 = -++$, $\mu_6 = -+--+$, $\mu_7 = -+---$, $\mu_8 = --+$, $\mu_9 = ----$.



$\mathcal{C}_4$

$\mathcal{C}_4$ is not uniform - nor composed uniformily of a smaller code. But it has the property that by inverting its words we found again a unitary code and, indeed, its symmetric image (symmetric in respect of the N.S. line !)

Since ergodicity is somewhat synonimous of irreversibility of time, we are put on the alert by this oddity.

Indeed, absorption is linked very closely with the problem of reading "backward" messages with an inverted code, but, without entering this amusing theory, we can see at once that $\mathcal{C}_4$ and all its family are not ergodic.

If a code is unitary the only sequences which let π_1 invariant are the complere messages, whose set is M'. In symbols, this means :

$\mu_1 \mu_2 \in M'$ and $\mu_1 \in M'$ implies $\mu_2 \in M'$

Suppose now that the same property be true on the other direction, i.e. that we had :

$\mu_1 \mu_2 \in M'$ and $\mu_2 \in M'$ implies $\mu_1 \in M'$

Let μ_1 be a complete message which is the unperturbated beginning of the transmission; μ'_1, its noise corrupted form and μ_2 any other complete message. By the above condition $\mu'_1 \mu_2$ may have a final scansion like that of $\mu_1 \mu_2$ if and only if μ'_1 is a complete message, too.

As this is usually not the case the error will go on till the end.

Codes which are unitary for both directions of time (anagrammatic codes) are not yet fully explored but a construction for various infinite families of them is known. With binary alphabet, there is just the one given above and its symmetric for less than 16 words. It is conjectured that there is still no more than 38 other one below 32 words (on about 10^{16} distinct usual unitary neat codes of this size or less!).

So the family is really exceptionally interesting and deserves further studies since with the uniform and the uniformily composed codes, anagrammatic codes are the only length-bounded codes escaping ergodicity.

References.

1. A.Sardinas and Patterson (1953) Convention records of the I.R.E.
2. B.Mandelbrot .
   1953. Proc.Symp.Comm.Theory. London.
   1954. Proc. Symp. Inf. Network.
   1955. Proc. Symp. Comm. Theory . London.
3. P. Dubreil.
   1941. Mem. Acad. Sci. p. 1-52.
   1951. Rendiconti di Math. 81 p. 289 - 306.
   1953. Bull. Soc. Math. (10) p. 183 - 200.
4. P.Elias.
   1955. Proc. Symp. Comm. Theory. London.
5. A.E.Laemmel.
   1953. ibid.

Part II

Appendix 0 : Some notations for semi group concepts :

A being any semi group, capital letters will represent subsets of it, small letters being reserved to elements of A . It will be understood that :

$XY = Z$ means that Z is the set of all elements $z = xy$ in A obtained as product of an $x \in X$ by an $y \in Y$.

Residuals :

x and y being two elements of A , the notation $x^{-1} y$ (resp. $y x^{-1}$) denotes the (eventually vacuous) set of all those elements z which satisfy $x z = y$ (resp. $z x = y$)

z is called a "residual". Various notations for it are to be found in the literature. The one below seems the most logical :

X and Y being two sets :

$X^{(-1)} Y$ is the set Z of all z such as $xz \in Y$ for at least one $x \in X$.

$X^{[-1]} Y$ is the set of all z such as $xz \in Y$ for all $x \in X$.

The same holds mutatis mutandis for $Y X^{(-1)}$ and $Y X^{[-1]}$.

Unitary and neat sub semi group.

These two fundamental notions are due to P. Dubreil (Mem. Ac. Sci. t. 63. 1941. p. 1-52 and Bull. Soc. Math. 81. 1953 p. 289-306).

X is a sub semi group of A if and only if :

$$X^2 = XX \subset X.$$

X is unitary on the left (right) if and only if :

$xy \in X$ and $x \in X$ ($y \in X$) implies $y \in X$ ($x \in X$)

In symbols : $X^{(-1)} X \subset X$. (on the right: $XX^{(-1)} \subset X$).

X is neat on the right (left) if for all x there is at least one y with $xy \in X$ ($yx \in X$).

In symbols : $X A^{(-1)} \supset A$ (on the left $A^{(-1)} X \supset A$).

Appendix I. Generalized Sardinas and Patterson's condition.

Let M be the free semi group generated by the letters $\{M_\alpha\}$; Q its sub semi group generated by the words $M_i \in Q_0$, these being for the sake of the demonstration any finite strings of letters.

Theorem : A n.a.s.c. for Q to be isomorphic with a free semi group is that for any three elements : x, y, z

$x \in Q$; $z \in Q$; $xy \in Q$; $yz \in Q$ imply : $y \in Q$.

In symbols : $$Q^{(-1)} Q \cap Q Q^{(-1)} \subset Q \qquad \text{I.1}$$

Proof :

Let $u = a_1 a_2 \dots a_{n-1}$ be a finite string of letters pertaining to Q . Write $u_{ii'}$ as an abbrevation for the subsequence $a_i a_{i+1} \dots a_{i'-1}$ obtained from u by amputation.

Call "critical indice" of u any indice i such as u_{1i} and u_{in} pertain both to Q and let Ju be the set of the critical indices of u .

We prove first :

If i and $i' \in Ju$ and $j \in Ju_{ii'}$ then $j \in Ju$.

Indeed, by hypothesis, u_{1i} , u_{ij} , $u_{ji'}$ and $u_{i'n}$, pertain all to Q so that the same is true of $u_{1j} = u_{1i} u_{ij}$ and $u_{jn} = u_{ji'} u_{i'n}$.

Now :

A n.a.s.c. for Q to be a free semi group is that $i, i' \in Ju$ entails $u_{ii'} \in Q$

Let $Ju = \{ 1 = i_1 < i_2 < \dots < i_m = n-1 \}$

The condition is sufficient for : (1) by the above lemna, the $u_{i_k i_{k+1}}$'s cannot be broken further and thence are words; (2) there is no alternative scansion of u , since this would imply the existance of a critical indice i' not contained in Ju.

The condition is necessary. Suppose $u_{ii'} \notin Q$.

Since j = 1 and j' = n fulfill (1) and (2) below, there exits $j, j' \in Ju$ such as:

(1) $j < i < i' < j'$

(2) u_{ji} , $u_{ji'}$, $u_{ij'}$, $u_{jj'}$ are all in Q

(3) $|j' - j|$ be a minimum.

$c = u_{jj'}$ admits at least two different scansions: one by chopping of $u_{ji'}$ and $u_{i'j'}$ into words; the other one chopping of u_{ji} and $u_{ij'}$. These are different for if u_{ji} and $u_{ji'}$ - for instance-had a common complete message $u_{jj''}$, as left divisor, the couple (j'', J') would satisfy (1) and (2) above and (j', j') would not be minimal.

The demonstration is finished for it is enough now to take $x = u_{ji}$, $y = u_{ii'}$ and $z = u_{ij'}$ for obtaining the terms of the theorem.

Remark I. If Q is (left or right) unitary it corresponds automatically to a code for $Q Q^{(-1)} \subset Q$ for instance implies obviously $\mathcal{I}(1)$.

Infinite families of codes which are not unitary neither on the left nor on the right may be constructed but no example still is known in which all words have a bounded length.

Remark II. Refinement of Sardinas and Patterson's method leads to an important result which we do not prove here :

A n.a.s.c. for the existence of a fixed finite number $L < \infty$ such that the knowledge of the $m + L$ first letters of a message allows always an unambiguous decoding of the first m letters whatever be m is that the code be bounded and unitary on the left.

Appendix II. Syntactic equivalence and related concepts.

The notion of syntactic equivalence had been already met in 1951 (C.R.Acad.Sci. 232 p.1987 - 1989) by M. Teissier working on abstract semigroups, but this concept does not seem to have still received the attention it deserves.

Actually, we shall state the main results for a slightly broader relation :

" $\succ$ (K) " defined in any semi group A in respect of any subset $K \neq \emptyset$ of A.

II.1. $a \succ b$ (K), if and only if, whenever $xby \in K$ then : $xay \in K$ ..

In symbols :

$a \succ b$ (K) $\rightleftarrows \forall_A x, y,\ xby \in K \Rightarrow xay \in K$.

We prove, now, the following algebraic properties of $\succ$

II.2. $a \succ a$ (K) (obvious).

II.3. $a \succ b$ (K) and $b \succ c$ (K) entail: $a \succ c$ (K) (obvious)

II.4. $a \succ b$ (K) and $c \succ d$ (K) entail: $ac \succ bd$ (K) (if II.1. is true for all x , $y \in A$, it is still true for all $x \in Au$ and $y \in vK$; thus, owing to associativity :

$a \succ b$ (K) entails : $uav \succ ubv$ (K), whatever be $u, v \in A$.

In particular, if $c \succ d$ (K), one has :

$ac \succ bc$ (K) and $bc \succ bd$ (K)

i.e. : $ac \succ bd$ (K), in view of II.3).

II.5. If ρ is any relation on A satisfying II.2, II.3. and II.4. and the further condition : $a \rho b$ and $b \in K$ entail: $a \in K$ (for short " K is upper saturated for ρ "), then: $a \rho b$ entails : $a \succ b$ (K). (First, $a \succ b$ (K) and $b \in K$ entail : $a \in K$: it is enough to make $x = y = \emptyset$ in II.1.

Now, if ρ is as described, $a \rho b$ implies : $xay \rho xby$ and, by the supplementary condition, whenever : $xby \in K$, then : $xay \in K$ so that, $a \rho b$ entails $a \succ b$ (K)).

We define now :

II.6. $a \equiv b$ (K), if and only if, in the same time $a \succ b$ (K) and $b \succ a$ (K)

In view of II.3,4,5,6 :

(II.2)' $a \equiv a$ (K)

(II.2)" $a \equiv b$ (K) implies: $b \equiv a$ (K)

(II.3)' $a \equiv b$ (K) and $b \equiv c$ (K) imply: $a \equiv c$ (K)

(II.4)' $a \equiv b$ (K) and $c \equiv d$ (K) imply: $ac \equiv bd$ (K)

(II.5)' $a \equiv b$ (K) and $a \in K$ imply: $b \in K$

and " $\equiv$ (K) " is maximal with these properties.

Not every congruence on A is surely a syntactic equivalence but at least we can state:

II.7. If ρ is any congruence on A with classes K_1 , K_2 , ... , K_α , then, $a \rho b$ is equivalent to : $a \equiv b$ (K_1) and $a \equiv b$ (K_2) and... $a \equiv b$ (K_α). (By the very definition of ρ , $a \rho b$ entails $xay \rho xby$ for all x, $y \in A$, that is to say : $xay \in K_i \rightleftarrows xby \in K_i$

Conversely suppose a and b not in relation ρ ; then $a \in K_i$ and $b \in K_{i'}$, with $i \neq i'$ and accordingly $a \not\equiv b$ (K_i).

The problem of determining in a usefull way just how many and which of the K_i's are enough for reconstructing in this way a given ρ is still open.

The meaning of $\equiv$ will be clearer if we consider the special case where A is a finite group.

II.8. If A is a finite group, $a \succ b$ (K) implies : $b \succ a$ (K) and is equivalent to : $a^{-1} b \in G = \bigcap_{x \in A} x^{-1} K x$ = the largest normal subgroup contained in K . Thus, $a \equiv b$ (K) is equivalent to : $a \equiv b$ (G).

(If A is a finite group $Ku \subset K$ entails $Ku = K$, whatever be $u \in A$. Further, if ρ is any congruence, there is a normal subgroup G_ρ of A such that all saturated $K's$ (i.e. all subset

of A satisfying : $a \in K'$ and $a \rho b$ entails $b \in K'$) may be put under the form : $K' = \bigcup_{k' \in K'} k' G_\rho$.

Now II.2 may be written : $K\, y^{-1} a^{-1} b\, x \subset K$ for all $x, y \in A$ so that $a \gtrdot b$ (K) entails $a \equiv b$ (K). Owing to II.5, there is G, normal, with $K = \bigcup_{k \in K} k G$. From the above special form of II.2, it comes in particular that $a^{-1} b \in G$ implies : $a \equiv b$ (K) and, by the maximal property of $\equiv$(K), that $G = \bigcap_{x \in A} x^{-1} K\, x$).

Fondamental semi groups.

Call φ the semi group homomorphism associated to $\equiv$ (K).

II.9. $a \gtrdot b$ (K) in A is equivalent to :

$\varphi a \gtrdot \varphi b$ (φK) in φA.

(On one hand : $xby \in K$ implies $\varphi xby = \varphi x\ \varphi b\ \varphi y \in \varphi K$. On the other hand, since K is upper saturated for $\gtrdot$(K), $\varphi x\ \varphi b\ \varphi y \in \varphi K$ entails $xby \in K$).

II.10. In φA, $\varphi a \equiv \varphi b$ (φK), if and only if : $\varphi a = \varphi b$. (Since by II.9 $\varphi a \equiv \varphi b$ (φ K) implies $a \equiv b$ (K) i.e. $\varphi a = \varphi b$ in φ A). So, once applied φ to the original A, iteration of the process in respect of φK does not give any further not trivial homomorphism, or, as one could say : φK is syntactically simple in φA.

We recall that if ψ is any homomorphism and σ the congruence defined by : $x \sigma y$ if and only if $\psi x = \psi y$, the fact that a complex U be saturated for σ is equivalent to : $\psi^{-1} \psi U \subset U$.

This again is equivalent to :

For all x : $\psi(U\, x^{-1}) = \psi U\, (\psi x)^{-1}$.

($a \in U\, x^{-1}$ means $ax \in U$, which entails $\psi a\ \psi x \in \psi U$ i.e. $\psi a \in \psi U\, (\psi x)^{-1}$; conversely, $\psi a\ \psi x = \psi q \in \psi Q$ entails $ax = q \in Q$; this is the very definition of saturation).

The following result allows the building of all semi groups (A $\supset$ K) such as (φ A $\supset \psi$ K) be isomorphic with a given $\overline{A} \supset \overline{K}$ where $\overline{K}$ is syntactically simple in $\overline{A}$.

II.11. Be given $A \supset K$ and an homomorphism ψ.

A n.a.s.c. for $\varphi_K = \varphi_{\psi K} \circ \psi$ is that $\psi^{-1} \psi K \subset K$. (Necessary : if $a \in A - K$, $k \in K$ and $\psi k = \psi a = \overline{k} \in \psi K$, then : $(\varphi_{\psi K} \circ \psi) k = (\varphi_{\psi K} \circ \psi) a$ and $\varphi_K a \neq \varphi_K k$.

Sufficient : if $a \equiv b$ (K) entails $a \sigma b$, then it entails: $\psi a \equiv \psi b$ (ψK).).

Now, we prove that φ preserves features important for the translation theory :

II.12. Let $Q = K \supset K^2$ be a sub semi group from A. Any one of the following properties is true in the same time in A and in φA :

$Q^{-1} K \cap Q Q^{(-1)} \subset Q$; $Q^{(-1)} Q \subset Q$; $A \subset Q\, A^{(-1)}$

(obvious due to the saturation property and the fact $X\, Y^{(-1)}$ is just a notation for $\bigcup_{y \in Y} X y^{-1}$).

Prefixes.

Definition : we call right prefixes the equivalence classes for the relation :

$a \sim b$, if and only if : $a^{-1} K = b^{-1} K$. ($\sim$ is the "principal equivalence" of P. Dubreil)

II.13. $a \sim b$ entails: $ax \sim bx$ for all x and: $a \equiv b$ (K) is equivalent to: $ya \sim yb$ for all $y \in A$.

(Classical: $au \in K \Rightarrow bu \in K$ for all u, entails in particular :

$axv \in K \Rightarrow bxv \in K$ for all v , by specialising $u = xv$.

Now, $a \equiv b$ (K) could just as well be written : $(ax)^{-1} K = (bx)^{-1} K$ for all x, or : $K\, (ya)^{-1} = K\, (yb)^{-1}$ for all y).

As an immediate consequence :

II.14. The representation A^* of the $x \in A$ as applications of the set $\prod^*$ of the right prefixes into itself is an isomorphic representation of $\overline{A} = \varphi$A.

(a, b $\in \pi_i$ entails $ax, bx \in \pi_i \supset \pi_i x$. On the other hand $\pi_i x \sim \pi_j y$ for all $\pi_i \in \prod^*$, entails $x \equiv y$ (K) and reciprocally.)

Of course "left prefixes" do exist just as well and play some role in the theory. In general, their set ${}^*\prod$ is quite different from $\prod^*$ and, accordingly, the two matrix representations of φA as translations on the right or on the left - although isomorphic - are not in general equivalent. In connection with these points it may be useful, at times, to visualise in a somewhat different way the notion of prefix :

Let κ be the relation between subsets of A defined by $X \kappa Y \Rightarrow XY \subset K$

II.15. The operations $X \to K\, (X^{(-1)} K)^{(-1)} = X^*$ and $X \to (K\, X^{(-1)})^{(-1)} K = {}^*X$ are the Galois closures corresponding to κ and the closed sets of the type X^* (or ${}^*X$) are the right (left) prefixes.

(Indeed, by definition, the right prefix π_i^* containing x is the set of all x' such as whenever $xy \in K$, then $x'y \in K$ too; if making $X = x$, we obtain exactly $\pi_i = X^*$. The link with "Galois closure" concepts is straight forward (see for instance : Birkhoff; Lattice Theory, chap. IV)).

This remark leads very easily to the following useful proposition :

II.16. Let A be a free semi group generated by the finite set of letters $\{A_o\}$ and $K = M$ be a sub semi group of A generated by M_o.

If M is isomorphic to a free semi group and if every word in the generating set M_o has a length bounded by $L < \infty$, then φA is finite. (Owing to Theorem I.1. $ax \in M$ implies that the first critical indice of ax at the right of the end of d be separated from this point by less than L letters.

Further $ax \in M$ implies that $axm \in M$ for all $m \in M$. So, the set of all x such as $ax \in M$ for a given a has the form $M'M$ where M' is a subset from the bounded set of all right divisors of the elements in M_o. It follows that there is only a finite number n of right prefixes and from II.I5 a finite number, too, $\leq 2^n$ of right prefixes. The proposition is then a direct consequence of II.I4.)

Finally if dropping the boundedness condition for M_o we add that M be unitary on the left, we obtain a partial result which will be used in the next appendix.

II.I7. Let $\approx$ denote the transitivity closure of the relation $\widetilde{\approx}_o$ defined by : $a \widetilde{\approx}_o b$ whenever exist $m, m' \in M \cup \emptyset$ and $c \in A$ with : $a = mc$ and $b = m'c$.

Under the above specified conditions : $a \approx b$ entails $a \sim b$ and each $\approx$-class $\widetilde{\pi}_i$ contains a single well defined element c_i admitting no $m \in M$ as left divisor.

(By the definition, $a \approx_o ma$ so that one may cancel as many as initial $m \in M$ as to obtain a c which cannot be left simplified in this way. Obviously no two such c and c' can be in relation $\approx_o$.

If one had $c \approx c'$, this would mean that for a family $\{a_i\}$ there exists the relations :

$c \approx_o a_1$; $a_1 \approx_o a_2$; ..... ; $a_\alpha \approx_o c'$.

But $c \approx_o a_1$ and $a_1 \approx_o a_2$ imply :

$a_1 = mc$ and $a_1 = m_1 c_1$ and $a_2 = m_2 c_2$

with $m_i \in M$ and c_i left-simplified as above.

Owing to the fact that M is isomorphic to a free semi group, this implies :

$c_1 = c$ and $m_1 = m$ i.e. : $c \approx_o a_2$.

Finally, by the unitary character of M, $ax = mcx \in M$ entails : $cx \in M$ and $bx = m'cx \in M$ i.e. : $a \sim b$).

Actually, $a \approx b$ means exactly that, apart for an initial complete message, a and b correspond to the same node on the coding tree. A broader generalisation of the notion of syntactic equivalence is skeched in App. V.

Appendix III. Super coding :

The results in this section are not optimal. For the sake of simplicity we state the propositions under a form which is valid in the case of neat unitary coding. Some of them would have to be radically altered in the broader set-up alluded to at the end of App. II.

A is any semi group and $K = M$ a sub semi group of A, unitary on the left, neat on the right (U.N.).

Thus - (from II.1) - M is isomorphic by θ to a free semi group B. Let L be U.N. in B. This isomorphism sends L onto a subset $\theta L = N \subset A$ and we have :

III.1. N is a U.N. subgroup of A contained in M.

(That $N^2 \subset N \subset M$ is obvious. Owing to the isomorphism θ from $L \subset B$ to $N \subset M$, if $ax \in N$ then $\theta^{-1}(ax) \in L$ and if, further, $a \in N$, then $\theta^{-1} a \in L$ so that L unitary in B (i.e. $\theta^{-1} x \in L$) implies N unitary in A. In the same way N is neat in A if and only if $L \subset B$ and $M \subset A$ are both neat).

III.2. Any class $\widetilde{\pi}_i^N$ for $\approx (N)$ may be represented as the logical product of a $\approx (L)$ class, π_j^L , and a $\approx (M)$ class, $\widetilde{\pi}_k^M$.

(Let $\widetilde{\pi}_i^N$ have c for minimal representative - as defined in II.I6.- There is a unique decomposition of $c = m_1 m_2 \ldots m_\alpha c'$ with $m_i \in M$ and c' : a minimal representative of $\widetilde{\pi}_k^M$. On the other hand : $m = m_1 m_2 \ldots m_\alpha$ is, too, a minimal representative of π_j^L owing to the isophormism.)

III.3. If $a \equiv b$ (N), then : $a \equiv b$ (M)

(In view of II.I4 and II.I6 it suffices to prove that $a \approx b$ (N) implies $a \sim b$ (M). Let $c = m_1 m_2 \ldots m_\alpha c'$ be the minimal representative for the common $\approx (N)$ class $\widetilde{\pi}_i^N$ of a and b. Any $d \in \widetilde{\pi}_k^M$ has the form $d = mc$ with $m \in M$ so that $d^{-1} M$ depends only on c). It is an immediate consequence of III.3. that if φ_N and φ_M are the syntactic homomorphisms attached to N and M :

III.4. $\varphi_M A$ is an homomorphic image of $\varphi_N A$.

It must be observed that this is not generally true when N is any sub semi group contained in M in contrast with what happens when A is a group. The only perfectly general result is the far less interesting :

III.5. If K and K' are two subsets of A :

$a \gtrdot b$ (K) and $a \gtrdot b$ (K') entails : $a \gtrdot b$ $(K \cap K')$

($xby \in K \cap K'$ entails $xby \in K$ and $xby \in K'$

The last relations entail $xay \in K$ and $xay \in K'$

i.e. $xay \in K \cap K'$).

For filtering, one needs the very simple :

III.6. M being U.N. in A, there is for any pair of bounded $x, y \in A$ at least one bounded N M which is U.N. in A and such as: $x \not\equiv y$ (N).

(The only case of interest is $x \equiv y$ (M) :

Let $x = m_1 m_2 \ldots m_\alpha c$; $m'_1 m'_2 \ldots m'_\alpha c' = y$

($m_i \in M$) and $c\,d \in M$. Take for $N \subset M$ any U.N. code admitting $x\,d \in N_o$ and not $y\,d \in N_o$).

Appendix IV. The main ergodic theorem.

Let $A \supset Q$ be any two semi groups with $(-1)_Q$ $Q \cap Q$ $Q^{(-1)} \subset Q$ as in II.1.

Definition:

A will be said absorbing for Q if whatever a , b $\in$ A, there is at least a bounded u- depending eventually on a and b - such as : $a\,u \sim b\,u$ (Q).

Observe first that the problem is only interesting if Q is neat on the right in A. If not, there is indeed a trivial prefix π_o into which any x is sure to fall sooner or later ; this being characterized by :

$x \in \pi_o$ entails: for no u, $x\,u \in Q$.

Thus we are led to include the condition that Q be neat on the right in A. It will be very convenient to restrict even the problem to the case where :

1) A has a finite basis A_o.
2) Q is strongly neat - i.e. any bounded $x \in A$ admits at least one bounded right residual $u \in x^{-1}Q$.

By having read "backwards" the current paper, we know that all this implies that Q be unitary on the left. If not we look for a proof at appendix VI.

V. 1. A n. a. s. c. for A to be absorbing is that for all $x \in A$ there exists at least one bounded $q \in Q$ with $x\,q \in Q$ (i.e. $q \in x^{-1}Q$).
(Sufficient : Let $x\,y \in Q$ and $y\,q' \in Q$; Q being unitary : $x\,q \sim y\,q'$.
Necessary : Let $x \in Q$ and then $v \in u^{-1}Q$ then: $Q \supset x\,u\,v \sim y\,u\,v$ and $u\,v \in Q$.)

IV.2. If the conditions of IV.1. are fulfilled, there is for any finite set $X = \{x_1, x_2, \ldots\ldots x_\alpha\}$ a common bounded absorbing u :
(i.e. satisfying: $x_i\,u \in Q$ for all $x_i \in X$).
(Take first : $q_1 \in x_1^{-1}Q$ ($\neq \emptyset$ by hypothesis) then: $q_2 \in (x_1 q_1)^{-1}Q \neq \emptyset$ etc.
$u = q_1\,q_2 \ldots\ldots q_\alpha$)

IV.3. A sufficient condition for A not to be absorbing is that Q pertains to one of the three following families :

1) Uniform codes
2) Uniformly composed codes.
3) Anagrammatic codes (i.e. with $Q\,Q^{(-1)} \subset Q$) or to any coding obtained from these by U. N. supercoding.

(1) Q is the set of all sequences whose length is a multiple of $k \neq 1$ so that $x \notin Q$ entails : $x\,q \notin Q$ for all $q \in Q$.

(2) Q is the set of all sequences made of $k \neq 1$ words from P_o where $P \supset Q$ is U.N. If $x \in P - Q$, so is it of every $x\,q$ with $q \in Q$.

(3) Q unitary on the right implies that: $x\,q \in Q$ and $q \in Q$ only if $x \in Q$).

IV.4. If Q admits a bounded basis the IV.3 condition is necessary. (Proof involves the full theory of Suschkevitsch's Kern Gruppe plus some appendices. However, let it be sketched roughly.)

$\varphi\, A = \bar{A}$, finite, admits a minimal sub demi group $\mathcal{G}$ such as $x\mathcal{G}y \subset \mathcal{G}$ for all $x, y \in \bar{A}$. $\mathcal{G}$ is the minimal bilateral ideal of $\bar{A}$. $\mathcal{G}$ can be built out of a group $\mathcal{g}$ by some, not too immediate, method. (Suschkevitsch . Math. Ann. 99. 1928 . p.30-50. Cf, too: Clifford Am. J. Math. 71, 1949, p. 835) The condition for $\bar{A}$ to be absorbing is :
1) That $\mathcal{g}$ be reduced to a single element ε.
2) That $\mathcal{G}$ be the unique minimal left ideal of $\bar{A}$. (i.e. $\mathcal{G} = D$ where D is the minimal set with $xD \subset D$ for all x).

The way the elements of $\bar{A}$ operate on those of $\mathcal{G}$ can be represented by matrices (Schutzenberger, C. R. Acad. Sci. 18/VI/1956) in an isomorphic way when A is syntactically simple. Provided that Q is not unitary on the right (family (3)), if II is violated, there is a first homomorphism, which reduces $\bar{A}$ to $\bar{A}_1$ satisfying it.

If I is still violated, there exists again a second homomorphism sending $\bar{A}_1$ into $\bar{A}_e$ satisfying I and II. The inverse image of $\bar{Q}_1$ for the first homomorphism is $P \supset Q$ which is still unitary on the left (family 2) unless we be dealing with a member of family 1.)

Appendix V . Introduction of probabilities.

Let A and Λ be two d.s.l.g. with generating sets (or "alphabets") A_o and Λ_o. Let it be given a probability distribution $\mathcal{P}$ on Λ_o: $P = Pr(\lambda_{oi})$ for $\lambda_{oi} \in \Lambda_o$ and suppose that a source be generating sequences of λ_{oi}'s , independently, and according to $\mathcal{P}$. A coding machine transforms each of the successive λ_{oi} 's into a string of letters $q_{oi} = \theta\lambda_{oi}$ in A and a receiver is observing continuously this sequence at the end of the communication set-up. Whatever be the sub-set $Q_o = \{q_{oi}\}$ (generating a sub semi group $Q \subset A$), we call the stochastic process observed by the receiver the stochastic discrete semi group language (A_o , Q_o, $\mathcal{P}$).

In what follows it will be systematically assumed that the transmission has started at the time finitely remote in the past and that the receiver has followed the process from its very first letter. In fact, our discussion deals only with what could be called a "Prefix theory" and not with a full syntactactic theory - which we have no space to develop here. So the notation Pr (a x | a) will denote the probability for the receiver having observed the initial string of letters "a" to see it later followed by "x". However in view of their simplicity and general interest - we introduce first some concepts broader than it is strictly necessary here.

Definition : $(a , b) = (b , a) = \text{l. u. b.}_{x \in \Lambda} \left| \log Pr(a\,x \mid a) - \text{Log } P_r(b\,x \mid b) \right|$
v.0.1. $(a , a) = 0$; $0 \leq (a , b) \leq (a , c) + (c , b)$.

.0.2. : whatever be c : $(a\,c\,,\ b\,c) \leqslant 2\,(a\,,\ b)$.
(if: $\Pr(a\,c\,x) \neq o$; $\Pr(a\,c\,x \mid a\,c) = \dfrac{\Pr(a\,c\,x \mid a)}{\Pr(a\,c \mid a)}$

Definition: The classes on A for the right regular relation $a \overset{\mathcal{L}}{\sim} b \geqslant (a, b) = o$ will be called the stochastic prefixes.
($\mathcal{L}$ is right regular since $(a, b) = o$ entails $(ax, bx) = o$ for all $x \in A$)

The stochastic prefix to which pertains any a may be considered as a sufficient statistic or "resume exhaustif" in the sens of G. Darmois, of the past of the message and does not need for its definition that of a s.d.s.g.l. Relations stronger than $\overset{\mathcal{L}}{\sim}$ should be considered for such problems as Wiener's times series prediction.

V.1. A n.a.s.c. for $q \overset{\mathcal{P}}{\sim} q'$ for all $q\,,\ q' \in Q$, is that Q be unitary on the left and that Θ be a coding.
(To each a we associate the set $a \cap Q$ of its right multiples which are in the time members of Q.

We say that $q \subset a\,A \cap Q$ is minimal if for all decomposition of q under the form $q = q'q''$ $(q\,,q'' \in Q)$

$a = q'\,a'$ with $a' \neq \phi$.
Let the set of all such minimal $q \subset a\,A \cap Q$, be called the right envelope of d. By definition : $\Pr(a) = \sum_{\lambda \in \Theta^{-1} E^{\circ}_{q}} \Pr(\lambda)$
(Θ^{-1} is eventually a one-to-many operation in our perfectly general set up). If, and only if, Q is unitary on the left, E°_{q} reduces to q itself when $q \in Q$. In this case we have $\Pr(q\,c \mid q) = \Pr(c)$ whatever be $c \in A$, so that $\Pr(qc \mid q) = \Pr(c \mid q')$ i.e. :
$(q\,,\ q') =$ when $q\,,\ q' \in Q$.

If Q is not unitary or if Θ does not give a coding for at least a couple $q' = q\,r$ with $q,\ q' \in Q$, $r \notin Q$, $\Pr(q\,r \mid q) \neq \Pr(r)$ and $(q\,,\ q') \neq o$).

If comparing with chapter 12 of Feller's Probability Theory, one sees that the case of coding with Q unitary on the left coincidates rigorously with the assumption that the scansion symbol "/" is a recurrent event. This identity has been observed and its consequences developed first by B. Mandelbrot (C.F. C.G. Proc. Symp. Inf. Networks. 1954 p. 203) allowing by this connection an enlargement of the theory towards the continuous case which we have systematically avoided until now.

It will be observed that, in the unitary case stochastic prefixes correspond to nodes on the coding tree and that what we did with our definition was just reducing the stochastic process under consideration to the canonical form of a Markoff chain. It must be underlined that the number of $\mathcal{L}$ classes is in general infinite when Q is not unitary on the left even if all words are length bounded.

Appendix VI. Admissibility of U.N. coding.

In what follows we assume that the generating set A_o of A contains a finite number of elements.
As in App. V. we consider a D.s.g.l. Λ generated by Λ_o and any homomorphic application Θ of Λ onto a sub semi group $Q \subset A$; $|a|$ where $a \in A$ denotes its length.

Definitions : $N'(\ell) = \sum_{\lambda \in \Lambda} \delta(\ell, |\Theta\lambda|)$ with $\delta(\ell, |\Theta\lambda|)$
$= 1$ if $\ell = |\Theta\lambda|$
$= o$ if $\ell \neq |\Theta\lambda|$
($N(\ell)$ is the number of sequences in A of length ℓ which pertains to Q, each being weighted according to the number of its Q decompositions.)

$$h(t) = 1 - \sum_{\lambda_i \in \Lambda} t^{|\Theta\lambda_i|}$$

h (t) is the "structure function" of B. Mandelbrot.
We recall first the well known result :

VI.1. The generating function $H(t)$ of the $N(\ell)$ is $1/h(t)$ and N (N) may be expressed by :

$$N'(\ell) = \sum \rho_i^{-\ell} \beta_i(\ell)$$

where $\{\rho_i\}$ is the set of the zeroes of h (t) and $\beta_i(\ell)$ is a polymial which reduces to a constant β_i if and only if ρ_i is simple.
In view of its importance we single out ρ the zero of smallest modulus of h (t) . Owing to the expression of h (t) : $o \leq \rho < \infty$ and ρ is a simple root.

VI.2. A necessary condition for Θ^{-1} to give a coding (i.e. $Q^{(n)} Q \cap Q Q^{(n)} \subset Q$) is that $k^{-1} \leqslant \rho$
(If not, for large enough ℓ, $N(\ell) \geqslant k^{\ell}$; so that at least one $q \in Q = \Theta\lambda$ corresponds to two different λ 's.

VI.2. A necessary condition for Q to be strongly neat on the right is that $\rho \leqslant k^{-1}$
(If Q is strongly neat, to every one of the k^{ℓ} sequences of length ℓ in A corresponds at least one sequence in Q of length $\ell' = \ell + L$ where $L < \infty$. So for large enough ℓ :
$N(\ell) + N(\ell+1) + \ldots N(\ell+L) \geqslant k^{\ell}$

VI.3 To any structure function h (t) with $\rho \geqslant k^{-1}$ corresponds at least one Q which is unitary on the left.
($h(k^{-1}) \geqslant 0$ (Szilard's inequality). If the equality sign does not hold, develop $h(k^{-1})$ in ascending powers of k and consider it as minus the value for $t = k^{-1}$ of a new function $h'(t)$.
$\overline{h}(t) = h(t) = h'(t)$ is again a structure function for which $\overline{\rho} = k^{-1}$.
Write $\overline{h}(t) = (1-kt)(1+\eta_1 t + \eta_2 t^2 + \ldots)$
The η_i's satisfy $o \leq \eta_i \leq k\eta_{i-1}$. It is easy to check that these inequalities imply that the η_i's may be the numbers of not terminal nodes at length i from the apex of the coding tree corresponding to $\overline{Q}$ unitary and neat. The tree associated with the original h (t) is easily recovered by pruning from it the nodes and words introduced by $h'(t)$. Q is still unitary but no more neat.)

Consequences:

VI.5. The class of the U.N. codes is a complete admissible class and the structure function of an admissible code admits the root K^{-1} . (The construction of a P for which a U. N. code is optimal is well known. On the other hand if $\rho = K^{-1}$ - i.e. if the tree T does not correspond to a neat Q , it is easily seen how one at least of the words may be replaced by a shorter one)

Remark : A parallel line of argument could have been based on the function:
$h^*(t) = 1 - \sum_{\lambda_{oi} \in \lambda_o} \Pr(\lambda_{oi})\, t^{|\theta \lambda_{oi}|}$
deriving from the determinant of the transition matrice in the underlying Markoff process. Along this line, admissibility is proved (B. Mandelbrot) by using the Hartley - Shannon's information : $\sum_i p_i \log p_i$

VI.6. A n.a.s.c. for Q to be admissible is that for no a : u a v $\notin Q$ for all length bounded u , v $\in A$.
(Necessity is obvious since the existence of such an a would make $\rho > K^{-1}$ Sufficiency derives directly from the same argument as in VI.2.)

VI.7. If Q contains " a " as in VI.6., one may add to Q at least a word of the form x a - (or : a x) - without destroying the coding character of Q .
(Proof is combinatorial and is based on the refinement of Sardinas Patterson's algorithm.)

VI.8 If Q satisfies $Q \subset Q^{(-1)}$ $Q \cap Q$ $Q^{(-1)}$ and is admissible, a n.a.s.c. for it to be strongly neat on the right is that it be unitary on the left.

(The initial condition imposes $\rho = K^{-1}$. If Q , neat on the right, has Q_o for generating set and if $Q_o' \subset Q_o$ is the set of all words which are not proper right multiple of another word, then Q' generated by Q' is still neat on the right and is unitary on the left by definition. Since h (t) and h' (t) admit both the minimal root $\rho = K^{-1}$ they are identical and $Q_o' = Q_o$. If Q is unitary on the left, Q" the minimal U.N. code containing Q is, again for the same reason, identical with Q).

THE LOGIC THEORY MACHINE A COMPLEX INFORMATION PROCESSING SYSTEM

Allen Newell and Herbert A. Simon[1]
The RAND Corporation, Santa Monica, Calif.
and the Carnegie Institute of Technology,
Pittsburgh, Pa.

Abstract

In this paper we describe a complex information processing system, which we call the logic theory machine, that is capable of discovering proofs for theorems in symbolic logic. This system, in contrast to the systematic algorithms that are ordinarily employed in computation, relies heavily on heuristic methods similar to those that have been observed in human problem solving activity. The specification is written in a formal language, of the nature of a pseudo-code, that is suitable for coding for digital computers. However, the present paper is concerned exclusively with specification of the system, and not with its realization in a computer.

The logic theory machine is part of a program of research to understand complex information processing systems by specifying and synthesizing a substantial variety of such systems for empirical study.

Introduction

In this paper we shall report some results of a research program directed toward the analysis and understanding of complex information processing systems. The concept of an information processing system is already fairly clear and will be made precise in Section I, below. The term 'complex' is not so easily disposed of; but it is the crucial distinguishing characteristic of the class of systems with which we are concerned.

We may identify certain characteristics of a system that make it complex:

1. There is a large number of different kinds of processes, all of which are important, although not necessarily essential, to the performance of the total system;

2. The uses of the processes are not fixed and invariable, but are highly contingent upon the outcomes of previous processes and on information received from the environment;

3. The same processes are used in many different contexts to accomplish similar functions towards different ends, and this often results in organizations of processes that are hierarchical, iterative, and recursive in nature.

Complexity is to be distinguished sharply from amount of processing. Most current computing programs for high speed digital computers would not be classified as complex according to the above criteria, even though they may involve a vast amount of processing. In general they call for the systematic use of a small number of relatively simple subroutines that are only slightly dependent on conditions. In order to distinguish such systematic computational processes from the processes we regard as complex, we shall call the former algorithms, the latter heuristic methods. The appropriateness of these terms will become clearer as we proceed.

One tactic for exploring the domain of complex systems is to synthesize some and study their structure and behavior empirically. This paper provides an explicit specification for a particular complex information processing system --a system that is capable of discovering proofs for theorems in elementary symbolic logic. We will call the system the logic theorist (LT), and the language in which it is specified the logic language (LL). This system is of interest for a number of reasons. First, it satisfies the criteria of complexity we have listed above. Second, it is not so large but that it can be hand simulated (barely). Third, the tasks it can perform are well-known human problem solving tasks--it is a genuine problem solving system. Fourth, there are available algorithms, and a realization of at least one of these algorithms (the Kalin-Burkhart machine),[2] that can perform these same tasks; hence, the logic theorist provides a contrast between algorithmic and heuristic approaches in performing the same problem solving tasks.

The task of this paper, then, is to specify LT with sufficient rigor to establish precisely the complete set of processes involved and exactly how they interact. This is a lengthy and somewhat arduous undertaking but one that the authors feel is required in the present state of knowledge. As a result, the paper largely abstains both from comment on the more general significance of the ideas and techniques introduced, and from relating these to contemporary work.[3]

The plan of the paper is to give, in Section I, a description of the language, LL, in which LT

---

[1] The authors are indebted to Mr. J.C.Shaw of the RAND Corporation, who has been their partner in many aspects of this enterprise and particularly in undertaking to realize the logic theorist in a computer--work that will be reported in subsequent papers.

[2] See B. V. Bowden[1]

[3] We should like to make general acknowledgment of our indebtedness for many of the ideas incorporated in LL and LT to two areas of vigorous contemporary research activity: (1) to research on automatic programming of digital computers, for the approach to the construction of LL; and (2) to research on human problem solving, for the basic structure of the program of LT. In addition we should like to record a specific indebtedness to the work of O.G.Selfridge and G.P.Dinneen on pattern recognition, which clarified many basic conceptual issues in the specification and realization of complex information processes.

will be specified. In Section II there is given a verbal description of LT, which is closely enough tied in to the formal program to motivate most of the latter. Finally, in Section III, the program is given in full detail.

I

Language for Information Processing Systems

The two major technical problems that have to be solved in studying information processing systems by means of synthesis may be called the specification problem and the realization problem. To study all but the simplest of such systems, it is necessary to make a complete and precise statement of their characteristics. This statement, or specification, must be sufficiently complete to determine the behavior of the system once the initial and boundary conditions are given. An example, familiar to mathematicians, is a system specified by n first order differential equations in n variables.

Once the specification has been given, a second problem is to find or construct a physical system that will behave in the manner specified. This can be a trivial or an insurmountable task. For example, it is relatively easy to find electrical circuitry that will behave like a system of linear differential equations; it is rather difficult to represent by circuitry most kinds of nonlinear systems. We will call the problem of finding or constructing the physical system the realization problem, and the particular physical system that is used the realization.[4]

Although this paper is concerned exclusively with the specification problem, the form of language chosen is dictated also by the requirements of realization. Since an important technique for studying the behavior of complex systems is to realize them and to study their time paths empirically under a range of initial and boundary conditions, they must be specified in terms that make this realization relatively easy.

The high speed digital computer is a physical system that can realize almost any information processing system and our research is oriented toward using it. Its limitations are in overall speed and memory, rather than in the complexity of the processes it can realize. The machine code of the computer is the language in which a system must ultimately be specified if it is to be realized by a computer. Conversely, however, once the system is correctly specified in machine code, the realization problem is essentially solved; for the computer can accept these specifications, and will behave like the system specified.

The machine code, although suitable for communicating with the computer, is not at all suitable for human thinking or communication about complex systems. For these purposes, we need a language that is more comprehensible (to humans), but one that can still be interpreted by the computer by means of a suitable program. Technically, such a language is known as a pseudo-code or interpretive language. Hence the two problems of specification and realization of an information processing system are subsumed under the single task of describing the system in an appropriate pseudo-code.

This paper is concerned solely with specifying the system of LT. The particular language, LL exhibited here has not been coded for a computer. However, one very similar to it, which is less convenient for exposition, is in the process of being coded and will be the subject of later papers. Here, no further mention will be made of the relation of the logic language to computers.

The terms of the language that are undefined--its primitives--determine implicitly a set of information processes that are to be regarded as elementary and not reducible, within the language, to simpler processes. The more complex processes are to be specified by suitable combinations of these elementary processes. Generally speaking, the elementary processes in LL are of the nature of information processes: that is, their inputs and outputs are comprised of symbolized information.

Information Processing Systems: Basic Terms

An information processing system, IPS, consists of a set of memories and a set of information processes, IP's. The memories form the inputs and outputs for the information processes. A memory is a place that holds information over time in the form of symbols. The symbols function as information entirely by virtue of their capacity for making the IP's act differentially. The IP's are, mathematically speaking, functions from the input memories and their contents to the symbols in the output memories. The set of elementary IP's is defined explicitly, and through these definitions all relevant characteristics of symbols and memories are specified.

Particular systems can be constructed from the memories and processes of an IPS that behave in a determinate way once the initial information in the memories is given (initial conditions), along with whatever external information is stored in the memories during the course of the system's operation (boundary conditions). Each such particular system we call a program, IPP. Thus an IPS defines a whole class of particular IPP's, and conversely, an IPP consists of an IPS together with a set of rules that determines when the several information processes will occur. The logic language is an IPS; the logic theorist is an IPP. Many variations of LT could be constructed with the same IPS.

Symbolic Logic

The logic language handles information referring to expressions in the sentential calculus and their properties. This paper assumes some

[4] We prefer "realization" to "simulation," for the latter implies that what is being imitated is another physical system. Since the specification is an abstract set of characteristics, not a physical system, it is not correct to speak of "simulating" the specification.

familiarity with elementary symbolic logic,[5] and only a resume of the notation will be given.

The sentential calculus deals with variables, p, q, ..., A, B, ..., a, b, ..., which are usually interpreted to mean sentences. These variables are combined into expressions by means of connectives. The primitive connectives of Whitehead and Russell (and ours) are "not" (-) and "or" (v). In this paper we shall have occasion to use only one other connective: "implies" (→), which is defined by:[6]

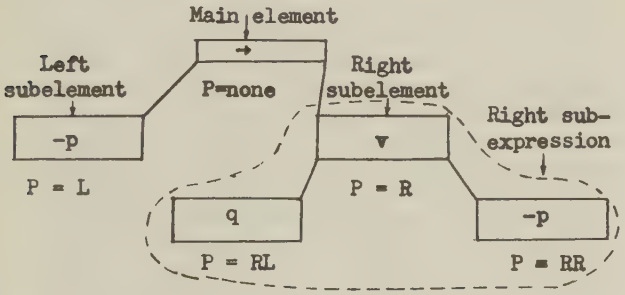
1.01 $p \rightarrow q \ =_{def} -p \vee q$ (Read: (p implies q) is equivalent by definition to (not-p or q).)

Coding

A logic expression, X, is represented in the IPS by a set of elements, E, one corresponding to each variable and to each connective (excluding the punctuation dots and negation symbols) in the logic expression. Each element holds a number of symbols that refer to the various properties of the element. (Note that the term "element" and not the term "symbol" is used in this paper to refer to the variables and connectives in logic expressions. Symbols denote properties of elements, and to each element there correspond a number of symbols.) An example will show what is meant by these terms.[7] Consider the expression 1.7:

1.7 -p .→. q v -p ((not-p) implies (q or not-p))

The entire sequence is the expression, X(1.7). It consists of the elements -p, →, q, v, -p. The expression may be written in "tree" form, as follows, where the rectangles indicate the elements:



[5]For definiteness, we have used the system of A.N. Whitehead and Bertrand Russell.[3] An introduction sufficient for our purposes will be found in D. Hilbert and W.Ackermann.[2]

[6]For ease of reference, we shall use the numbers employed by Whitehead and Russell to identify particular propositions and definitions, only omitting the asterisk (*) that they insert in front of the number.

[7] We follow Whitehead and Russell in using dots in place of parentheses as punctuation. It is unnecessary here to give exact rules for numbers of punctuation dots.

The main connective at the top is called the main element, EM (1.7). The other elements are reached through a series of Left and Right branches from the main element. With each element there is associated a subexpression, namely, the subtree of which that element is the top element.

The symbols in each element provide the following information, which will be explained more fully as we proceed.

Symbol

- G The number of negation signs (-) before the expression. In the figure above, two elements--those containing the variable p--have G = 1; all the rest have G = 0. If a negation applies to a whole expression it appears in the element associated with that expression.
- V Whether the element is a variable or not.
- F Whether the element is free, i.e., available for substitution. This is relevant only if E is a variable.
- C The connective (v or→). This is relevant only if E is not a variable.
- N The name of the variable or expression. In X(1.7), there are variables named "p" and "q".
- P The position of the element in the tree. This is represented by a sequence of L's and R's, counting branches from the main element. In the figure, the P for each element is shown beneath the element.
- A The location of the whole expression (not the element) in storage memory.
- U Whether the element is to be viewed as a unit or not. The term "unit" will be explained later.

The eight symbols defined above characterize completely each element and the expression in which it occurs. For many purposes, however, it is convenient to define additional symbols ("descriptive symbols") that correspond to interesting or important properties of expressions. In LL, three such descriptive symbols, represented as small positive integers, are defined. These are:

- H The number of variable places in an expression. Thus X(1.7) has three variable places: P = L, RL, and RR; hence, H(1.7) = 3.
- J The number of distinct variables (i.e., distinct names) in the expression, ignoring negation signs. Since X(1.7) contains the names "p" and "q", J(1.7) = 2.
- K The number of levels in the expression. The number of levels corresponds to one plus the maximum number of letters in P for any element in the expression. Hence, K(1.7) = 3.

Memory Structure

There are two kinds of memories, working memories and storage memories. The major distinction --that all information to be processed must be

brought in from the storage memories to the working memories and then returned--will be brought out clearly when we define the elementary IP's. Structurally, the working memories hold single elements, E, with additional spaces for the symbols H, J, and K. Hence, we can picture a working memory unit as:

The storage memories consist of lists. A list holds either a whole logic expression or some set of elements generated during a process, such as a set of elements having certain properties. Each list of logic expressions has a location, symbolized by A. The elements are placed in the list in arbitrary order, since the information in each element is sufficient to locate it unequivocally in the tree of the logic expression. (The ordering of the list is used only to carry out searches.) For example, X(1.7) might be listed in the storage memory thus:

No limitations are imposed here on number of memories, either working or storage. In actual fact, the number used is not large.

Three particular lists have special locations in storage memory that can be referred to directly in IP's: (1) the theorem list, T, of all axioms and theorems that have previously been proved; (2) the active problem list, P; and (3) the inactive problem list, Q. Each list consists of the main elements of the appropriate expressions (theorems or problems, respectively) in arbitrary order. For the rest, the storage memory is entirely unspecialized.

Information Processes

A term that specifies an IP is called an instruction, by analogy with computer terminology. As Figure 1 shows, an instruction consists

Fig. 1

of an operation part, three reference places (left, L, center, C, and right, R), and a branch location, B. The kinds of operations that can be performed by an IPS will depend, first, on what elementary IP's are postulated, and second, on what restrictions are placed on how they can be combined. For the moment, the exact nature of the elementary processes is unimportant; for concreteness, the reader may think of the following as typical: transferring information from memory x to memory y, or adding the number in memory x to the number in memory y.

The reference places refer to the working memories, so that the same operation may operate on different memories at different times and under different circumstances. The working memories will be designated by small integers, 1, 2, ..., and by the letters x, y, z.

No direct reference is made in an instruction to any storage memory, except T, P, and Q. Lists are located by the A stored within elements belonging to the lists; and elements within a list are located by their relation to known elements. An example will make this clear. A typical operation involving the storage memory is:

OPER	L	C	R	B
FR	x		y	

which reads: Find the element that is the right subelement of E(x)--i.e., of the element in working memory x--and put it in working memory y. The operation is executed thus: Working memory x contains the A(x) that is the location of the expression in which E(x) occurs. Memory x also contains the symbol P(x). Since we wish to put in y the right subelement of E(x), P(y) is by definition obtained by appending an R to P(x). Hence, we can determine P(y), and locate E(y) by going to storage memory A(x) and searching the list of its elements in order until we find the element with the correct P. We then transfer this element, which is the one we want, to working memory y.

Programs and Routines

The rule of combination of IP's is simple: any one IP may follow another. We shall consider time to be discrete, using it essentially as an index, and shall assume that only one process occurs at a time. We say that a particular IP has control when it is occurring. Thus, when a sequence of IP's occurs one after the other in consecutive time intervals, there occurs a series of transfers of control from each IP to the next in the sequence.

The operation of any IP includes a processing component and a control component. The processing component changes the memory content of the IPS; the control component transfers control to another IP. In some IP's, processing is the significant component. In these the transfer of control is independent of the memory contents at the time the IP occurs. In certain other IP's, control is the significant component. These do not alter memory contents, but transfer control to various IP's depending on the memory contents when they occur. In other IP's both processing and control components are significant.

Control. We allow only a binary branch in control at any one instruction. Normally, control passes in a linear sequence through a set of IP's. We write this sequence vertically. Each instruction is considered to have a location in the sequence. For branch instructions (those in which the control component transfers control to one of two IP's depending on memory content), control transfers either (1) to the next instruction in the sequence or (2) to the instruction named in the branch location, B. These locations are designated by letters A, B, C, ... In

Figure 2, Instruction #1 transfers controls to #2; #2 transfers control to #3 or branches to A (which is #4) depending on memory content; #3 transfers control to #4; #4 transfers control to #5 or branches back to B, which is #1.

Each control operation can be reversed in sense by putting a minus sign in front of the operation name. The effect of the minus sign is simply to reverse the condition of transfer. That is, if CC-A transfers to A when two specified numbers are equal, then -CC-A transfers to A when these numbers are unequal.

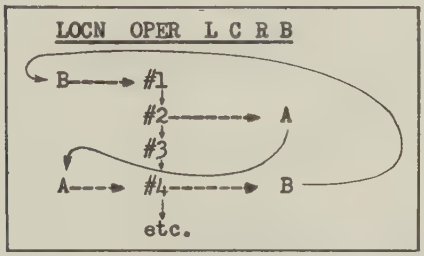


Fig. 2

Routines. We will call such a list of instructions with a control network a routine, again, in direct analogy to computer terminology. Notice that a routine satisfies our definition of a program (IPP): if all the memories referred to have specified initial contents, the routine determines their contents at all later times covered by its duration.

If we postulate a set of elementary information processes, each specified by an instruction, it might be supposed that each routine would define a new (non-elementary) information process. This is not the case, for in LL the format of an instruction (Figure 1) allows reference to not more than three working memories and to not more than one branch. Hence, only those routines may be regarded as definitions of IP's which satisfy the following conditions:

1. The routine contains branches to not more than two instructions outside the routine;
2. Not more than three working memories that are to be referred to subsequently are changed by the routine. This means that even though other working memories are changed, there is no way to refer to these memories in subsequent routines.

Within these restrictions we can define a set of new IP's in terms of the elementary IP's, then another set of IP's in terms of both the elementary and defined IP's, and so on; thus creating a whole hierarchy of IP's and their corresponding routines. The elementary IP's and the hierarchy of defined IP's for LT are given in Section III, and its structure as explained in some detail in Section II.

The restrictions imposed above on numbers of branches and working memories in IP's have the following two consequences for the structure of the routines that are used to define IP's:

1. A working memory can be used only within the routine in which it is introduced. That is, working memories introduced in a particular routine cannot be referred to when control is in any other routine, except as noted in rule 2. For this reason, no ambiguity arises from using the same names, 1, 2, . . ., for different memories in distinct routines.

2. Within the routine that defines a particular IP, reference may be made to the working memories that are designated in the reference places of that IP. Let I_1 be an instruction that appears in the routine defining I_2. The symbols L, C, R in I_1 refer, respectively, to the working memories in the left, center, and right reference places of instruction I_2, in whose definition I_1 occurs. (See, for example, the first instruction, FEF, in the routine given in full at the end of the section.) Some such arrangement is obviously required if the defining routine is to have any connection with the instruction it defines.

Elementary Processes

In LT there are forty four different elementary processes. These represent variations on eight types of operations. The remainder of this section will be devoted to a description of these types, and an enumeration of the elementary processes that belong to each type. Separate, explicit definitions for each elementary IP are given in Section III. The first letter in the name of an operation designates the type to which it belongs: A for assign, B for branch, C for compare, F for find, N for numerical, P for put, S for store, and T for test.

Find instructions obtain information from storage memory on the basis of stated relationships, and put it in specified working memories. An example, FR-x-y (Find the right subelement of E(x) and put it in y), has already been described. Two other Find instructions are very similar: FL (Find the left subelement) and FM (Find the main element).

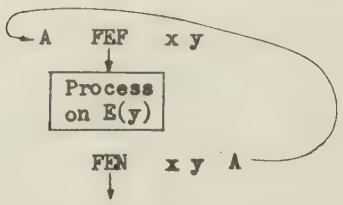
Other Find instructions involve the ordering relation on the lists. An example is:

OPER	L	C	R	B
FEF	x	y		A

This reads: Find the first element in X(x)—the expression associated with E(x)—in the list A(x), and put this element in y. Then go to next instruction, but if no element is found, branch to instruction A. Here the order of elements is essential since there may be many elements in X(x). This kind of operation is used to start a search, and it is always combined with an instruction, FEN, for continuing and terminating the search:

OPER	L	C	R	B
FEN	x	y		A

This reads: Find the element in X(x) that is next in order after E(y) and put it in y. When such an element is found, branch to A; if none is found, transfer control to the next instruction in sequence. FEF and FEN together allow the familiar cycling or iteration that is a common feature of computing routines:



(after all elements of X(x) have been processed)

The complete list of elementary Find instructions is:

FEF	FL	FM
FEN	FR	

Store instructions transfer information from working memory back to storage memory. An example is:

OPER	L	C	R	B
S		x		

This simply reads: Store E(x) in the storage memory. If the element in x is one that was previously withdrawn from storage, it will be replaced in its original location within A(x); if it is a new element in List A, it will be placed at the end of the list.

Another elementary Store instruction is SEN, which puts E(x) into storage memory at the end of the list A(y). A third is *SX, which simply stores a copy of X(x) in memory location A(y).[8]

The complete list of elementary store instructions is:

S	*SX	*SXL	*SXM
SEN	*SXE	*SXR	

Instructions belonging to the remaining six types are concerned only with working memory (See Figure 3). No complex processing may take place in storage memory, and conversely, as we have seen, no information may be stored in working memory except on a temporary basis.

Put instructions transfer information and symbols around the working memory. A typical Put instruction is:

OPER	L	C	R	B
PE		x	y	

This reads: Put E(x) in E(y). The operation leaves E(x) unchanged and duplicates it in E(y). The variations on this instruction correspond to the different symbols in an element that may need to be transferred. The list of Put instructions is:

PE	PCv	PU
PK	PC→	PUB

[8]Certain of the Store instructions are marked with an asterisk. These are treated as elementary operations in the present section and in Part I of Section III, but in Part II of Section III they are defined in terms of simpler operations.

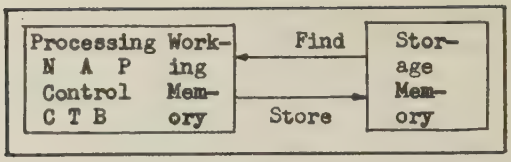


Fig. 3

Numerical instructions carry out the arithmetic operations. An example is:

OPER	L	C	R	B
NAG		x		

This reads: Add 1 to G(x). Operations are required to permit addition and subtraction for symbols G, H, J, K, and W. The list of Numerical instructions is:

NAG	NAH	NAJ	NSG
NAGG	NAK	NAW	NSGG

Assign instructions write in new names and locations in elements that are in working memory. One Assign instruction is:

OPER	L	C	R	B
AN		x		

This reads: Assign an unused name to E(x). The other Assign instruction, AA, assigns new list locations. There are, then only two Assign instructions:

AA	AN

Compare instructions belong to a class of pure control instructions. They compare two symbols for equality (or, if appropriate, for the relation "greater"); then transfer to the branch location if the condition is satisfied or to the next instruction in sequence if the condition is not satisfied. The sense of the branch on these and all other branch instructions can be reversed by a minus sign preceding the operation. A typical example is:

OPER	L	C	R	B
CC		x	y	A

This reads: If C(x) = C(y), branch control to location A; if not, go to the next instruction in sequence. That is, if the connective in x is identical with the connective of y, we branch to A. Notice that there is no change in memory content; only a transfer of control has occurred. The compare instructions are:

CC	CGG	CWG
CN	CKG	CPS

Test instructions are also control instructions. They test the properties of a single element, and transfer control accordingly. The variations of the type deal with different properties. An example is:

OPER	L	C	R	B
TU		x		A

This reads: If E(x) is a unit, transfer control to A; if not, go to the next instruction in sequence. TC→ transfers control if C(x) is "implies"; goes to the next instruction if C(x) is "or". The Test instructions are:

TV	TB	TU	TF
TC→	TN	TGG	

Branch instructions are unconditional control instructions that cause the program to branch to the indicated address instead of going to the next instruction in sequence. The simplest example is:

OPER	L	C	R	P
B				b

When this instruction is reached, the program simply branches to instruction b in the same routine.

When the instructions BHB or BHN occur in a routine, they cause the program to branch to an address determined by the higher-level instruction that the routine defines. For example, suppose BHB appears as one of the defining instructions within the routine:

OPER	L	C	R	B
MSb	x			b

Then, the occurrence of BHB will cause control to branch to the address b of MSb.

Suppose, further, that MSb appears as one of the instructions in the routine Ex, and that the instruction MDt appears immediately after MSb in Ex. Then, if BHN is one of the instructions in the routine MSb, its occurrence will cause control to branch to the next instruction after MSb in the higher routine, Ex, i.e., to MDt. Thus BHB and BHN are the instructions that terminate control by a particular routine, and cause control to transfer, respectively, to the branch designated in the higher-level instruction defined by the routine, or to the higher-level instruction that follows the routine. Instruction BHB produces the former transfer, BHN, the latter. The three Branch operations are:

B	BHB	BHN

Example. It will clarify matters and provide some introduction to the complete program given in Section III if we set forth in detail one of the simpler defined routines, the routine NH. This routine consists of six instructions, all of them primitives included in the list we have already given:

A	OPER	L	C	R	B
	NH	x			
	FEF	L	1		C
A	-CPS	1	L		B
	-TU	1			B
	NAH	L			
B	FEN	L	1		A
C	BHN				

Count the number of variable places in X(x), and record the result in H(x).

(1) FEF finds the first element in X(x) and puts it in working memory 1. If there is no element, it branches to C. (2) -CPS (note the negative sense) determines whether E(1) is a subelement of E(x). If it is not, control transfers to B; if it is, control transfers to the next instruction in sequence. (Henceforth we will abbreviate these transfers as →B and →next, respectively.) (3) -TU determines whether E(1) is a unit (i.e., is to be viewed as a variable). If it is not (negative sense), →B, if it is, →next. (4) NAH increases by 1 the number H(x). (Because of the previous branches, NAH will occur only if the element in 1 is viewed as a variable and is a subelement of the element in x.) (5) FEN finds the next element in X(x), puts it in working memory 1, and returns control to instruction A, whereupon the cycle is repeated from step (2). If there are no more elements, →next. (6) BHN terminates the routine after all elements in X(x) have been examined, and transfers control to the instruction that follows NH at the next higher level of the hierarchy of routines.

Conclusion

We have now completed our description of the language LL. We have outlined the coding system, the memory structure, the structure of the information processes, the routines, and the types of elementary processes. Further detail can be found by consulting Section III. In Section II we shall construct in this language a program, LT, that will permit the information processing system to solve problems in symbolic logic.

II

The Logic Theory Machine

In the language we have constructed, we have variables (atomic sentences): p, q, r, A, B, C, ... and connectives: - (not), v (or), → (implies). The connectives are used to combine the variables into expressions (molecular sentences). We have already considered one example of an expression:

1.7 $\quad -p \;.\!\rightarrow\!.\; q \vee -p$

The task set for LT will be to prove that certain expressions are theorems—that is, that they can be derived by application of specified rules of inference from a set of primitive sentences or axioms.

The two connectives, - and v, are taken as primitives. The third connective, →, is defined in terms of the other two, thus:

1.01 $\quad p \rightarrow q \;\text{ =def }\; -p \vee q$

The five axioms that are postulated to be true are:

1.2 $\quad p \vee p \;.\!\rightarrow\!.\; p$

1.3 $\quad p \;.\!\rightarrow\!.\; q \vee p$

1.4 $\quad p \vee q \;.\!\rightarrow\!.\; q \vee p$

1.5 $\quad p \;.\!\vee\!.\; q \vee r \;:\!\rightarrow\!:\; q \;.\!\vee\!.\; p \vee r$

1.6 $\quad p \rightarrow q \;.\!\rightarrow\!:\; r \vee p \;.\!\rightarrow\!.\; r \vee q$

Each of these axioms is stored as a list in the theorem memory, T, with all its variables marked free, F, in their respective elements.

From the axioms other true expressions can be derived as theorems. In the system of Principia Mathematica, there are two rules of inference by means of which new theorems can be derived from true expressions (theorems and axioms). These are:

Rule of Substitution: If A(p) is any true expression containing the variable p, and B any expression, then A(B) is also a true expression.

Rule of Detachment: If A is any true expression, and the expression A → B is also true, then B is a true expression.

To these two rules of inference is added the rule of replacement, which states that an expression may be replaced by its definition. In the present context, the only definition is 1.01, hence the rule of replacement permits any occurrence of (-pvq) in an expression to be replaced with (p→q), and any occurrence of (p→q) to be replaced with (-pvq).[9]

In this system, then, a proof is a sequence of expressions, the first of which are accepted as axioms or as theorems, and each of the remainder of which is obtained from one or two of the preceding by the operations of substitution, detachment, or replacement.

Example: prove 2.01, p → -p .→. -p:

(1) ⫶ p v p .→. p (axiom 1.2)[10]
(2) ⫶ -p v -p .→. -p (by subst. of -p for p)
(3) ⫶ p → -p .→. -p (by replacement on left)

The problem now is to specify a program for LT such that, when a problem is proposed in the form of a theorem to be proved (like 2.01 above), a proof will be discovered and constructed. First, it should be observed that there is a systematic algorithm for constructing such a proof should one exist. Starting with the five axioms, we construct all the theorems that can be obtained from them by a single application of the rules of substitution, detachment, or replacement.[11] We thus obtain the set of all theorems that can be obtained from the axioms by proofs not more than one step in length. Repeating this process with the enlarged set of theorems, we obtain the set of all theorems that can be derived from the axioms by proofs not more than two steps in length. Continuing, we finally obtain the set of theorems that can be derived by proofs not more than n steps in length.

Now, if the theorem in which we are interested possesses a proof k steps in length, we can, in principle, discover it by constructing all valid proof chains of length not more than k, and selecting any one of these that terminates in the theorem in question. This "in principle" possibility is in fact computationally infeasible because of the very large number of valid chains of length k that can be constructed, even when k is a number of moderate size. Under these circumstances, the rules of inference do not give us sufficient guidance to permit us to construct the proof we are seeking; and we need additional help from some system of heuristic.

The problem will be solved if we can devise a program for constructing chains of theorems, not at random, but in response to cues that make discovery of a proof probable within a reasonable computing time. For example, suppose the rules of inference were such as to permit any given proof chain to be continued, on the average, in ten different ways. Then there would be ten thousand proofs chains four steps in length (10^4). The expected number of proof chains that would have to be examined to find any particular proof by random search is five thousand. Suppose, however, that LT responded to cues that permitted eight of the ten continuations at each step to be eliminated from consideration. Then the number of proof chains four steps in length that would have to be examined in full would be only sixteen (2^4), and the expected number would be only eight.

The Program of LT

We wish now to describe the program of LT, which is given in full in Section III; hence, in the text we shall refer frequently to Section III for detail. We shall refer to each routine by its name (e.g., LMc for the matching routing), but we shall need some additional notation to refer to the main segments of routines that do not themselves have names. The names of these segments are given in Section III in the column marked "Seg." In each segment there is generally one main operation to be performed; and this main operation, or sub-routine, is usually surrounded by a number of procedural and control operations that fit it into the larger routine. In ordinary language, we would say that the "function" of the segment is to perform the main operation that is contained in it. For example, the main operation in the third segment of LMc is LSby, a substitution. The function of this segment in the matching program is to substitute one sub-expression for another in one of the expressions being matched. Hence, we will name the segment after the main operation: LMc(Sby). Similar designations will be used for the other segments of routines. This notation emphasizes the fact that each routine consists in a sequence (or branching tree) of main operations that are connected by procedural and test operations. Thus, an abbreviated description of the matching routine might be given as:

---

[9] As we shall see, 1.01 is not held in storage memory, but is represented, instead, by two routines for actually performing the replacements.

[10] The exclamation point in front of an expression indicates that the expression in question is asserted to be true. To designate an expression whose truth has not been demonstrated, we will use a question mark preceding the expression.

[11] A technical difficulty arises from the fact that there is an infinite number of valid substitutions. This difficulty can be removed rather easily, but the question is irrelevant for the purposes of this paper.

LMc

T Perform diagnostic tests
LMc Recursion of matching with next elements in logic expression
Sby Substitute the element y for the element x
Sbx Substitute the element x for the element y
CN Compare variables in x and y
Rp Replace connectives, if required and possible

The Substitution Method

Let us take as our first example the very simple expression, 2.01, for which we have already given a proof. We suppose that, when the problem is proposed, LT has in its theorem memory only the axioms, 1.2 to 1.6. We wish to construct a proof (the one given above, or any other valid proof) for 2.01.

As the simplest possibility, let us consider proofs that involve only the rules of substitution and replacement. We are now able to state the problem thus: how can we search for a proof of the expression by substitution without considering all the valid substitutions in the five axioms? We use two devices to focus the search. Both of these involve "working backward" from the expression we wish to prove--for by taking account of the characteristics of that expression, we can obtain cues as to the most promising lines to follow:

1. In attempting substitutions, we will limit ourselves to axioms (or other true theorems, if any have already been proved) that are in some sense "similar" in structure to the theorem to be proved. The routine that accomplishes this will be called the test similarity routine, CSm.
2. In selecting the particular substitutions to be made in a theorem that has been chosen for trial, we will attempt to match the variables in that theorem to the variables in the expression to be proved. Similarly, we will try to use the rule of replacement to match connectives. The routine in which these various operations occur is called the matching routine, LMc.

Using these devices, the proposed routine for proving theorems—the method of substitution, MSb—works as follows. MSb(Sm): search for an axiom or theorem that is similar to the expression to be proved. MSb(Mc): when one is found, try to match it with the expression to be proved; if a match is successful, the expression is proved; if the list of axioms and theorems is exhausted without producing a match, the method has failed. (Reference to Section III will show that there is another segment, MSb(NAW), that we have not mentioned. The function of this segment will be discussed later in connection with the executive routine.)

To see in detail how the method operates, we next examine the main operations, CSm and LMc, of the two segments of the substitution method. For concreteness, we will carry out these operations explicitly for the proof of the expression 2.01.

2.01 ? $p \rightarrow -p .\rightarrow. -p$

Test for Similarity, CSm. We must state what we mean by similarity. We start from a common-sense viewpoint and regard two propositions as similar if they "look" similar to the eye of a logician. In Section I we have already defined three characteristics of an expression that can be used as criteria of similarity. These are: K, the number of levels in the expression; J, the number of distinct variables in the expression; and H, the number of variables in the expression.[12]

Applying these definitions to 2.01 (routines NK, NJ, and NH, respectively), we find that K = 3, J = 1, and H = 3. That is, 2.01 has three levels, one distinct variable (p), and three variable places. We may write this:

$$D(2.01) = (3,1,3)$$

In the same way, we can write descriptions for the various sub-expressions contained in 2.01--in particular, the sub-expressions to the left and to the right of the main connective, respectively. We have for these:

$$DL(2.01) = (2,1,2); \text{ and } DR(2.01) = (1,1,1)$$

Now, we say that two expressions, x and y, are similar if they have identical left and right descriptions; i.e., if DL(x) = DL(y) and DR(x) = DR(y). The routine for determining whether two theorems are similar, CSm, consists of two segments: (1) CSm(D), a description segment, and (2) CSm(CD), a comparison of descriptions. The description segment is made up of four description routines, D, one each to compute DL(x), DR(x), DL(y), and DR(y). The comparison segment is made up of two compare description routines, CD, one of which compares DL(x) with DL(y), the other DR(x) with DR(y).

A diagram of the hierarchy of principal subroutines in testing similarity will look like this:

[12] The assertion is that two expressions having the same description "look alike" in some undefined sense; and hence if we are seeking to prove one of them as a theorem, while the other is an axiom or theorem already proved, then the latter is likely construction material for the proof of the former. Empirically, it turns out that with the particular definition of similarity introduced here, in proving the theorems of Chapter 2 of Principia Mathematica about one theorem in five that is stored in the theorem memory turns out to be similar to the expression we are seeking to prove. It is easy to suggest a number of alternative, and quite different criteria that would be equally symptomatic of "similarity." Uniqueness is of no account here; all we are concerned with is that we have some criteria that "work"— that select theorems suitable for matching.

In the case of 2.01, the segment MSb(Sm) will search the list of axioms and theorems and will find that axiom 1.2 is similar to 2.01:

1.2 ! p ∨ p .→. p

for it, too, has the descriptions: DL(1.2) = (2,1,2); DR(1.2) = (1,1,1). Moreover, 1.2 is the only axiom that has this description.

Matching Expressions, LMc. Next we carry out a point-by-point comparison between 2.01, the expression to be proved, and 1.2, the axiom that is similar to it. We start with the main connectives, and work systematically down the tree of the logic expressions--always as far as possible to the left. In the present case the order in which we will match is: main connective (P = none), connective of left sub-expression (P=L), left variable of sub-expression, (P=LL), right variable of sub-expression (P=LR), and right sub-expression (P=R).

The matching routine is fairly complicated, consisting of six segments, but not all segments are employed each time two elements are matched. The first segment, LMc(T), and the initial operations of most of the other segments consist of tests that determine whether the two elements to be matched are already identical, whether they can be made identical by substitution (if one is a free variable) or by replacement (if both are connectives), or--finally--whether matching is impossible. The second segment, LMc(LMc), is a recursion of the matching routine with each of the next lower pair of elements in the tree of the expression. This recursion segment operates only if the elements to be matched in LMc are identical connectives (or have been made so).

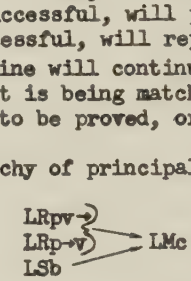
The third and fourth segments, LMc(Sby) and LMc(Sbx), apply the rule of substitution when the tests have shown this to be appropriate. LMc(Sby), which is executed whenever E(x) is a free variable,[13] simply substitutes the expression X(y) for E(x). LMc(Sbx), which is executed whenever E(y) is a free variable, substitutes the expression X(x) for E(y). In both cases, of course, substitution must take place throughout the whole expression in which the free variable occurs. This is taken care of automatically by the process LSb. Also, since LMc matches X(x) to X(y), LMc(Sby) has priority over LMc(Sbx), as a careful examination of the test network will reveal.

The fifth segment, LMc(CN), reports the successful termination of the matching program if E(x) and E(y) are identical variables, its failure if they cannot be made identical by substitution.

The sixth segment, LMc(Rp), operates when E(x) and E(y) have different connectives. The segment replaces the connective in x by the connective in y whenever this replacement is legitimate, and then returns control to the recursion segment.

By virtue of the recursion segment, the matching routine will attempt to match each pair of elements; if successful, will proceed to the next pair; if unsuccessful, will report failure. Hence, the routine will continue until it makes the theorem that is being matched identical with the expression to be proved, or until the matching fails.

The hierarchy of principal routines looks like this:



Returning to our specific example of the two similar expressions, 1.2 and 2.01, we carry out the matching routine as follows:

2.01 ? p → -p .→. -p
1.2 ! A ∨ A .→. A

(We use A instead of p in 1.2 to indicate that the variable is free (F).)

a. The main connectives agree: both are →.

b. Proceeding downward to the left, the connective is → in 2.01, but ∨ in 1.2. To change the ∨ to →, we must have (because of the definition, 1.01, a - before the left-hand A in 1.2. This we can obtain by making the substitution of -B for A in 1.2. Having carried out this substitution, and having then replaced (-B ∨ -B) with (B→ -B), we have the following situation:

2.01 ? p → -p .→. -p
1.2' ! B → -B .→. -B

c. Proceeding again to the left, we find B in 1.2', but p in 2.01. We therefore substitute p for B in 1.2', and now find (after recursion through the remaining two elements) that we have a complete match:

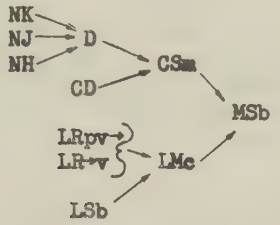
2.01 ? p→ -p .→. -p
1.2'' ! p→ -p .→. -p

Thus, we have discovered a proof of 2.01 (in fact, precisely the proof we gave before), which consists in substituting the variable -p for the variable in 1.2, and replacing the connective ∨ in 1.2 with →.

This completes our outline of the method of substitution as a routine for discovering proofs in symbolic logic. The method may be viewed as an information process that is composed of a considerable number of more elementary in-

[13] Essentially, a variable is free when no substitution has yet been made for it. After any substitution it is bound and no longer available for subsequent substitutions. As previously noted, all variables in expressions stored in the theorem memory are free.

formation processes arranged to operate in highly conditional sequences. Each of the main components--the test for similarity routine, and the matching routine--is made up, in turn, of subroutines. The test conditions that control the branchings of the sequences depend in a number of instances upon the outcomes of searches through the theorem memory. Hence, the method of substitution represents a complex information process in the sense in which we have defined the term. Combining the two diagrams depicted above, we can illustrate the hierarchy of the main operations that enter into the substitution method:



The method is a heuristic one, for it employs cues, based on the characteristics of the theorem to be proved, to limit the range of its search; it does not systematically enumerate all proofs. This use of cues represents a great saving in search, but carries the penalty that a proof may not in fact be found. The test of a heuristic is empirical: does it work?

Moreover, the cues that are used in the method are not without cost. For example, in order to limit matching attempts to "similar" theorems, theorems must be described and compared. The net saving in computing time, as compared with random search, is measured by the reduction in the number of theorems that have to be matched less the cost of carrying out the search and compare for similarity routines. Stated otherwise, cues are economical only if it is cheaper to obtain them than to obtain directly the information for which they serve as cues.

To be sure, we have found a proof for one proposition in Principia; but how general is the substitution method? On examination of the 67 propositions in Chapter 2 of Principia, it appears that some 21 can be proved by the method of substitution, including for example: 2.01, 2.02, 2.03, 2.04, 2.05, 2.10, 2.12, 2.21, 2.26, 2.27. The remaining propositions evidently require more powerful techniques of discovery and proof. It is evident, for instance, that we must employ the rule of detachment.

The Method of Detachment

We will describe next the method of detachment, MDt, which, as its name implies, incorporates the rule of detachment. The method, of course, is not synonymous with the rule, but includes also heuristic devices that select particular theorems to which the rule is applied.

Let us review the principle of logic that underlies the method. Suppose LT must prove that expression A is a theorem; and assume that there are in the theorem memory two theorems, B and B→A. Then, by application of the rule of detachment to B and B→A, A is derivable immediately.

We can generalize this procedure by combining matching (substitution and replacement) with detachment. Assume that the theorem memory contains B" and B'→A'; that A is obtainable from A' by matching; and that B' is obtainable from B" by matching. Then we can construct a proof of A as follows: (1) By matching with B", B' is a theorem. (2) Since B'→A' is also a theorem, it follows by detachment that A' is a theorem. (3) By matching with A', A is a theorem.

This settles the problem of constructing a valid proof by the method of detachment. From the standpoint of the discovery of a proof employing this method, the trick lies again in narrowing down the search for B'→A' and B", so that these do not have to be sought through a very large scale trial-and-error search and substitution program.

Structure of the Detachment Method. The basic structure of the detachment method is quite similar to that of the substitution method, for both methods utilize the same basic operations. The first two segments of the detachment method, MDt(SmV) and MDt(SmCt), carry out searches for similar expressions, in a way that will be indicated more precisely below. The next segment, MDt(Mc), carries out a matching of any expression so found with the theorem to be proved. If the matching is successful, a new problem is created by the segment MDt(F). This problem is then attacked, in the final segment, MDt(MSb), by the method of substitution.

Again, designate by A the expression to be proved. In MDt(SmV) we search the theorem memory for theorems whose right sides are similar (by the test, CSm, described previously) to the whole expression A. If we find such a theorem (call it T), we go to segment MDt(Mc), and apply the matching operation to the right side of T and to A. If we are successful in the matching, we find the left side of T, MDt(P); and seek to prove by the method of substitution that it is a theorem, MDt(MSb). For if the left side of T is a theorem and T is a theorem, then by detachment, the right side of T is a theorem. But A can be obtained from the right side of T by substitution, hence is a theorem. (Note that a check is made to see that T has → for a connective.)

Contraction. If the detachment method fails to find a proof in the manner just described, a new attempt is made by means of the second segment, MDt(SmCt), employing a different criterion of similarity from the one we have used thus far. If the theorem is similar, the method proceeds with the matching segment exactly as before.

To see what is involved in this generalized notion of similarity, let us consider two expressions, A and A', with different descriptions. If A has more levels and variable places than A', it is still possible that A is derivable from A' by substitution--specifically, by substituting appropriate molecular expressions for the variables of A. For example, take as A the expression:

2.06 ? p→q .→: q→r .→. p→r

for which we have DL(2.06) = (2,2,2), DR(2.06) = (3,3,4); and take as A' the expression:

A' ? a .→. b→c

for which we have DL(A') = (1,1,1), DR(A') = (2,2,2).

If in A' we substitute p→q for a, q→r for b, and p→r for c, we obtain 2.06. Operating in the reverse direction, if we contract 2.06 by making the inverse substitutions, we obtain A'. We can therefore refer to A' as "2.06 viewed as contracted."

Since the purpose in searching for similar theorems is to find appropriate materials to which to apply the matching routine, there is no reason why we should not use this more general notion of similarity if it proves effective in finding materials that are useful.

In general, what parts of an expression should be considered as units in the search for proofs is not a "given" for the problem solver. LT makes an explicit decision each time it looks for similar expressions as to what subexpressions will be taken as units. In contracting 2.06, a decision has been made that the elements p, q, and r are too small, and that more aggregative elements, e.g., (p→q) = a, should be perceived as units.

Examination of the routines for describing expressions (NH, NK, NJ) will reveal that these routines in fact count units rather than variables. Normally, the variables are the units used in description, for VV precedes CSm in every program except MDt. In the latter program, however, it is sometimes useful to view expressions as contracted, by means of VCt.

Example of Proof by Detachment. To illustrate the method of detachment, let us carry out explicitly the proof of 2.06:

2.06 ? p→q .→: q→r .→. p→r

The reader may verify that this theorem cannot be proved by substitution in the axioms and earlier theorems. Moreover the detachment method without contraction will also fail, for there is no theorem whose right side is similar to 2.06. However, we have already seen that when we contract 2.06, we obtain:

A' ? a .→. b → c

where p→q has been contracted to a, q→r to b, and p→r to c. We now have DL(A') = (1,1,1) and DR(A') = (2,2,2), descriptions that are identical with the descriptions of the sub-expressions of the right side of 2.04.

2.04 ! A →. B → C : → : B → . A → C
A' a → . b → c

Having selected 2.04 by use of the routine MDt(SmCt), we now proceed to match its right side with 2.06 in segment MDt(Mc):

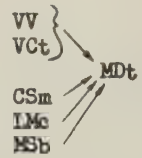
2.04 ! A .→. B → C :→: B .→. A → C
2.06 ? p→q .→: q→r.→.p→r
2.04' ! q→r .→: p→q.→.p→r .:→:. p→q .→: q→r.→.p→r

We have now created a new problem to replace the original one: to prove that the left side of 2.04' (the part underscored) is a theorem. We apply the method of substitution, MDt(MSb). The search of the theorem memory discloses 2.05 to be similar to the left side of 2.04', and we proceed to match them:

2.04'L ? q→r .→: p→q .→. p→r
2.05 ! A→B .→: C→A .→. C→B

It is easy to see that with the substitution of q for A, r for B, and p for C, the matching will be successful. Hence we have B (2.05 with the indicated substitution), and B→A (2.04'), from which A (2.06) follows by the rule of detachment.

The diagram below summarizes the principal routines incorporated in the method of detachment. A comparison of this diagram with the one for the substitution method shows clearly that both methods rest on the same component processes, with minor modifications and new combinations and conditions. The sole new process involved in detachment is the viewing of theorems as contracted.



The Chaining Method

A number of expressions that do not yield to the method of substitution can be proved by the method of detachment. We shall add an additional method, however, to the repertoire available to LT. We shall call this method chaining, MCh. Like the methods previously described, chaining involves heuristic procedures which we shall consider first.

Theorem 2.06, which we have just proved, embodies one form of the principle of the syllogism (2.05 is another form of this principle). Now suppose T_1, (p→q) is a true theorem, and T_2, (q→r) is another true theorem. Theorem 2.06 is of the form:

$$T_1 .\rightarrow. T_2 \rightarrow E$$

where E is (p→r), an expression not known to be true. By detachment, from ! T_1 and ! $T_1.\rightarrow.T_2\rightarrow E$, we get ! $T_2\rightarrow E$. By a second detachment, from ! T_2 and ! $T_2\rightarrow E$, we get ! E. Hence, if we know p→q and q→r to be true, we can construct a proof of p→r by means of two detachments with the use of 2.06. Instead of carrying through this derivation explicitly in each instance, we simply construct a program that makes direct use of the transitivity of syllogism. This proof method is the basis for chaining.

Suppose that we wish to prove A→C. We search for a theorem, T (with → for a connective) whose left side is similar to A, using the segment MCh(SmF). We match the left side of T with A, MCh(McF), and if we are successful, we have then proved a theorem of the form A→B, for T, as modified by matching, is of this form. We check first, in segment MCh(McR) whether we can simply match B to C. If we succeed, we have proved the theorem. If we fail, we now construct, by segment MCh(S), the expression B→C, and attempt to prove this expression by substitution, MCh(MSb). If we are successful, we now have a chain: A→B, B→C. Then by syllogism, as indicated above, we obtain A→C, the expression we wished to prove.

The procedure just described is chaining forward. Alternatively, we may chain backward. That is, to prove A→C, we may search for a theorem of the form B→C; then try to prove A→B by substitution.

Proof by the chaining method is illustrated by:

2.08 ? p → p

A search for theorems that have left sides similar to 2.08 yields 1.3, 2.02, and 2.07. The latter is:

2.07 ! p .→. pvp

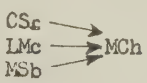
If we take 2.07 as the (A→B) of the schema given above, then B is (pvp). Two theorems have left sides similar to B: 1.2 and 2.01. An attempt to match the left side of 2.01 to the right side of 2.07 will be unsuccessful, but the matching is immediate with 1.2:

2.07 ! p .→. pvp
1.2 ! pvp .→. p

Hence we can take 1.2 as the (B→C) of the chaining method. We now form (A→C) by joining the left side of 2.07 to the right side of 1.2 by →. The result is 2.08:

2.08 ! p .→. p

The chaining method is summarized by the following diagram:



The Executive Routine

It remains to complete the specification of LT in two directions; first, to assemble the three methods that have been described into a coherent program; and second, to show how the information processes in terms of which LT has been described here can be specified precisely in terms of the elementary processes listed in Section I. The latter task is carried out in detail in Section III. We will turn our attention here to the former, which is embodied in the executive routine, Ex.

In its first segment, Ex(R), the executive routine reads a new expression that is presented to it for proof, and places it in a working memory.[14]

In the next three segments, Ex(MSb), Ex(MDt), and ExMCh), successive attempts are made to prove the expression by the methods of substitution, detachment, and chaining, respectively. If a proof is obtained by one of these methods, the executive routine writes the proof, Ex(WP); and stores the newly-proved theorem (changing all its variables to free variables) in the theorem memory, Ex(ST).

To explain what happens if the three methods are unsuccessful, we have to take up some details that were omitted above. These have to do with the creation of subsidiary problems and with stop rules.

Subsidiary problems. Both detachment and chaining are two-step methods. Suppose we wish to prove A. In detachment, we try to find a theorem, B→A, and if we are successful, we then try to prove B. The task of proving B we may call a subsidiary problem.

Suppose we wish to prove a→b. In chaining, we try to find a theorem, a→c, and if we are successful, we then try to prove c→b. The task of proving c→b is also a subsidiary problem.

Within both the detachment and chaining methods, only the method of substitution is applied to the subsidiary problem. If that method fails, failure is reported for the main problem. But before control is shifted back to the executive routine, the main element of the subsidiary problem is stored in the problem list, P, in the storage memory. (The operation that stores the problem in the problem list is the operation SEN that can be found in segment MDt(P) and segment MCh(P).)

When the three methods have failed for a given problem, the executive routine stores it in the inactive problem list, Q. It then selects from the problem list, P, an expression that is, in a certain sense, the simplest--specifically, an expression with the smallest possible number of levels, K, Ex(CK). It erases this new subsidiary problem from P; checks to make certain it does not duplicate one previously attempted, Ex(CX); and then tries to solve this subsidiary problem by the methods of detachment and chaining.[15] This sequence is repeated until some subsidiary problem is solved (in which case the main problem is also solved), or until no problems remain on the problem list, or until the other stop rule, to be described, comes into operation. In the latter two cases, the routine reports that it is unable to prove the theorem, Ex(WNP).

[14] Certain segments of Ex, in particular Ex(R), Ex(WP), Ex(ST) and Ex(WNP), are not written in Section III in terms of the primitives but are simply indicated by parentheses. It would be rather simple to formalize them, but this would further lengthen the description of the program.

[15] There is no need to attempt to prove the subsidiary problem by substitution, since an unsuccessful substitution attempt was made immediately before the expression was stored in the subsidiary problem list.

The check to prevent duplication of subsidiary problems, Ex(CX), is handled as follows: for each problem that is selected from list P by Ex(CK), a check is made, by Ex(CX), against all expressions in the inactive problem list, Q, and if the new problem duplicates any expression found there, it is dropped. The main operation of this segment, CX, applies the same basic tests of identity of elements that are applied in the matching program, but does not modify the expressions to make them match.

Stop Rules. Since all proof methods may fail, even if the expression given to LT is a genuine theorem, the executive routine needs a stop rule. One stop rule is provided by the exhaustion of list P, but there is no guarantee that the list will ever be exhausted. A second stop rule is provided by an operation that measures the total amount of "work" that has been done in attempting to prove a theorem, and that terminates the program with a "no proof" report when the total work exceeds a specified amount. The first operation in the substitution routine, NAW, tallies one for each time the routine is used. This tally is kept in a special location, W, in the storage memory. The executive routine, just before it seeks a new subsidiary problem, checks the cumulative tally in this register, Ex(CW), and if the tally exceeds a given limit, terminates the program. Since the substitution routine is used in each of the methods, the number of substitutions attempted seems to be one reasonable index of the amount of work that has been done.

This stop rule operates as a global constraint on the total work applied in trying to prove a single theorem. The rule does not govern the direction in which this effort is expended. The latter is determined by the priority rule previously described for selecting subsidiary problems from the problem memory and by the other elements of LT's program.

Learning Processes

The program we have described is primarily a performance program rather than a learning program. But, although the program of LT does not change as it accumulates experience in solving problems, learning does take place in one very important respect. The program stores the new theorems it proves, and these theorems are then available as building blocks for the proofs of subsequent theorems. Thus, in the theorems used as examples in this paper, 2.06 was proved with the aid of 2.05 and 2.04, and 2.08 was proved with the aid of 2.07. Without this form of learning it is doubtful whether the program would prove any but the first few theorems in Chapter 2 in a reasonable number of steps.

III

The Complete Program for the Logic Theorist

This Section is divided into two parts. The first part constitutes the program as described in the text, including the following routines: Ex; MCh, MDt, MSb; LMc, LSb, LRp→v, LRpv→, VV, VCt; CX; CSm, CD, D, NK, NH, NJ. These routines are preceded by a list of the most important primitive IP's—those that are used in several routines. Following each routine is a supplementary list of primitive IP's used in the definition of that routine.

The second part of this Section consists of routines for five IP's—those Store instructions that are marked with asterisks (*)—which up to this point have been treated as primitives.

Principal Primitive Instructions

A	OPER	L	C	R	B	
	B				b	Branch to b (→b).
	BHB					In higher instruction, →b.
	BHN					In higher instruction, →next.
	FEF	x		y	b	Find the first E in A(x) and put in y; if none, →b.
	FEN	x		y	b	Find the E in A(x) next after E(y), put in y; then →b. If none (end of list), →next.
	FL	x		y		Find EL(x) and put in y; if none, leave y blank.
	FR	x		y		Find ER(x) and put in y; if none, leave y blank.
	PE	x		y		Put E(x) in E(y);E(x) remains.
	S	x				Store E(x) back in A(x) (match on P); if not there, store E(x) at end of A(x).
	SEN	x		y		Store E(x) as next E in A(y); E(x) now last item in A(y).
	*SX	x		y		Store a copy of X(x) at (new) A(y). E(x) = M.
	TC	x			b	If C(x) = → (implies), →b.
	TV	x			b	If E(x) = V, →b.

A	OPER	L	C	R	B	Seg.	
	Ex						Executive routine
	(Read problem X)					R	
	(Put EM(X) in 1)						
	-MSb	1		C		MSb	
A	-MDt	1		G		MDt	
	-MCh	1		G		MCh	
	SEN	1	Q				X(1) is finished.
	CWG			H		CW	
B	FEF	P	1	H		CK	Find problem with lowest K.
	NK	1					
C	-FEN	P	2	D			
	NK	2					
	CKG	2	1	C			
	PE	2	1				
	PK	2	1				
	B			C			
D	E	1	P			CX	Remove duplicates of previous problems.
	FEF	Q	3	F			
E	CX	1	3	B			
	FEN	Q	3	E			
F	B			A			
G	(Write proof.)					WP	Succeeds in proving P.
	(X(1) a theorem)					ST	
	(Stop)						

```
H  (Write:no proof) WNP  Fails to find proof.
   (Stop)
```

Primitives

```
   CKG   x y   b    If K(x)>K(y), →b.
   CWG         b    If W (work done) > limit, →b.
   E     x y        Erase E(x) in A(y).
```

Note: There are six IP's in the executive routine that are not formally defined in LT. These are written in parentheses above: read problem, find problem and put in working memory 1, write proof, store expression as theorem, write "no proof", and stop.

```
A  OPER  L C R B    Seg.
                                Substitution method
                                 If can't prove X(x) by
   MSb   x      b                substitution, →b.

   NAW                 NAW  Count one unit of work.
   VV    L             Sm
   FEF   T 1   C
A  VV    1
   CSm   L 1   D
B  FEN   T 1   A             Find next T and repeat.
C  BHB

   SX    1 2           Mc
   LMc   2 L
   BHN
```

Primitives

```
   NAW                 Add one to W (work done).
```

```
A  OPER  L C R B    Seg.
                                Matching routine
                                 Match X(x) to X(y); if
   LMc   x y    b                can't, →b.

   CGG   C L   A   T
   CGG   L C   C        Now G(x) = G(y).
   TV    L     E
   TV    C     D
  -CC    L C   F
   FL    L 1       LMc
   FL    C 2
   LMc   1 2   H        Mc left subexpression.
   FR    L 3
   FR    C 4
   LMc   3 4   H        Mc right subexpression.
   BHN

A  TV    L     H   Sby
  -TF    L     H
B  NSGG  L C
   FM    L 5            Assures Sb everywhere.
   LSb   C L 5
   BHN

C  TV    C     H   Sbx
D -TF    C     H
   NSGG  C L
   FM    C 5            Assures Sb everywhere.
   LSb   L C 5
   BHN
```

```
E  TF    L     B   CN
  -TV    C     H
  -CN    L C   D
   BHN

F -LRp→v L     G   Rp   LRp's are self-testing.
   LRpv→ L     H
G  LMc   L C   H
   BHN
H  BHB
```

Primitives

```
   CC    x y   b    If C(x) = C(y), →b.
   CGG   x y   b    If G(x) > G(y), →b.
   CN    x y   b    If N(x) = N(y), →b.
   FM    x y        Find EM(x) and put in y.
   NSGG  x y        Subtract G(x) from G(y).
   TF    x     b    If E(x) is free, →b.
```

```
A  OPER  L C R B    Seg.
                                Substitution routine
                                 Substitute X(x) for
   LSb   x y z                   E(y) (=V) in X(z) (=M).

   FEF   L 1   F  F
A  CPS   1 1   B       E(1) must belong to X(x).
   CN    1 C   G
B  FEN   L 1   A
C  FEF   R 2   F  Sb   Search through X(z).
D -CN    2 C   E
   PE    L 3
   NAGG  2 3            G's add in Sb.
   SXE   3 2
E  FEN   R 2   D       Find next E(z), repeat.
F  BHN

G  AN    4         LSb
   LSb   4 C R
   B         C
```

Primitives

```
   AN    x              Assign an unused name to E(r).
   CN    x     b        If N(x) = N(y) →b.
   CPS   x y   b        If E(x) subelement of E(y)→b
                         (P(x) ⊃ P(y)).
   NAGG  x y            Add G(x) to G(y); result in
                         G(y).
  *SXE   x y            Store X(x) in A(y) in place
                         of E(y) (=V).
```

```
A  OPER  L C R B    Seg.
                                Detachment method
                                 If can't prove X(x) by
                                 detachment, →b. Store
   MDt   x      b                new problems in P.

   FEF   T 1   C  T
A  TC→   1     B       T must have C = →.
   VV    1
   FR    1 2
   VV    L         SmV
   CSm   L 2   D
   VCt   L         SmCt  Change view.
   CSm   L 2   D
B  FEN   T 1   A       Find next T and repeat.
C  BHB
```

```
D  SX    1 3          Copy, to work on T.
   FR    3 4
   LMc   4 L   B  Mc
   FL    3 5      P
   SXM   5 6        Create new X.
   S     6          Stored away fixed ME.
   SEN   6 P
   MSb   6     B  MSb
   BHN
```

Primitives

```
   *SXM  x y        Store X(x) at (new) A(y) as
                     main expression.
```

```
A  OPER  L C R B  Seg.
                        Chaining Method
                         If can't prove X(x) by
                         chaining, →b; Store new
   MCh   x     b         problems in P.
   -TC→  L     D  T     C(x) must be →.
   VV    L
   FL    L 1
   FR    L 2
   FEF   T 3   D
A  -TC→  3     C        T must have C = →.
   VV    3
   SX    3 4            Copy, to work on T.
   FL    4 5
   FR    4 6
   -CSm  1 5   B  SmF
   -LMc  5 1   E  McF
B  -CSm  2 6   C  SmB
   -LMc  6 2   F  McB
C  FEN   T 3   A        Find next T and repeat.
D  BHB
E  PE    2 5            Put E(2) and E(6) in
   PE    6 1             proper wkg. memory.
   -LMc  1 5   G  McR
F  AM    7      S       Create EM for new X.
   PC→   7              Fix connective.
   S     7              Store parts.
   SEN   7 P
   SXL   1 7
   SXR   5 7
   MSb   7     C  MSb
G  BHN
```

Primitives

```
   PC→   x          Put C(x) = → (implies).
  *SXL   x y        Store X(x) in A(y) as XL(y).
  *SXR   x y        Store X(x) in A(y) as XR(y).
```

```
A  OPER  L C R B  Seg.
                        Replacement of → with v.
                         If C(x) = →, replace
   LRp→v x     b         with v; if not →b.
   TC→   L     A  T
   BHB
A  PCv   L        Pv     Fix E(x).
   S     L
   FL    L 1              Fix EL(x).
   NAG   1
   S     1
   BHN
```

Primitives

```
   NAG   x          Add one to G(x).
   PCv   x          Put C(x) = v.
```

```
A  OPER  L C R B  Seg.
   VV    x                View variables as units.
   FEF   L 1      T
A  PUB   1                 Erase old unit.
  -TV    1      B
   PU    1      P
B  S     1
   FEN   L 1    A         Find next E and repeat.
   BHN
```

Primitives

```
   PU    x          Make E(x) a unit, (U).
   PUB   x          Make U(x) blank.
```

```
A  OPER  L C R B  Seg.
                        View as contracted
                         Make units of binary
                         expressions and
   VCt   x               isolated variables.
   TV    L      C  T
   FL    L 1       VCt
   FR    L 2
   TV    1      B
   VCt   1                 Recursion
   TV    2      E
A  VCt   2                 Recursion
   PUB   L
   S     L
   BHN

B  -TV   2      D
   PUB   1          Ct     Blank V's of Ct unit.
   S     1
   PUB   2
   S     2
   TN    L      C         Give X(x) a name if
   AN    L                 one needed.
C  PU    L
   S     L
   BHN

D  PU    1          VV     Make left (isolated)
   S     1                  variable a unit.
   B           A           XR(x) still to be done

E  PU    2
   S     2                 Make right (isolated)
   BHN                      variable a unit.
```

Primitives

```
   AN    x          Assign E(x) an unused name.
(See VV for PU and PUB)
   TN    x    b     If E(x) has a name →b.
```

```
A  OPER  L C R B  Seg.
                        Replacement of v with →.
                         If C(x)=v and G(EL(x))
                         >0, replace v with →;
```

```
   LRpv→ x      b          if not →b.

  -TCv   L      A  T
   FL    L 1
   TGG   1       C
  -TV    1      A
  -TSb   1      B
A  BHB

B  PE    1 2        Sb
   NAG   2
   FM    2 3
   LSb   2 1 3
   FL    L 1
C  PC→   L          P     Fix x.
   S     L
   NSG   1
   S     1
   BHN
```

Primitives

```
   FM    x y          Find EM(x) and put in y.
   NAG   x            Add one to G(x).
   NSG   x            Subtract one from G(x).
   PC→   x            Put C(x) = →.
   TGG   x     b      If G(x) >0 →b.
```

A OPER L C R B Seg.

Compare expressions

Compare X(x) with X(y); if they match, →b.

```
   CX    x y   b

   CGG   L C   B  T
   CGG   C L   B      G(L) = G(R), otherwise →B.
   TV    L     A
   TV    C     B
  -CC    L C   B      C(L) = C(R)
   FL    L 1     CX   Recursion down tree of
   FL    C 2          expressions.
  -CX    1 2   B
   FR    L 3
   FR    C 4
  -CX    3 4   B
   BHB

A -TV    C     B  CN  L and C both variables;
  -CN    L C   B      with identical names.
   BHB

B  BHN
```

Primitives

(For CC, CGG, and CN, see LMc)

A OPER L C R B Seg.

Similar expressions test

If DL(x) = DL(y) and DR(x) = DR(y), →b.

```
   CSm   x y   b

   FL    L 1       D
   FR    L 2
   D     1
   D     2
   FL    C 3
   FR    C 4
   D     3
   D     4
  -CD    1 3   A   CD
  -CD    2 4   A
   BHB

A  BHN
```

A OPER L C R B Seg.

Compare descriptions

If K(x) = K(y), J(x) = J(y), and H(x) = H(y) →b.

```
   CD    x y   b

  -CK    L C   A      Def: If K(x) = K(y) →b.
  -CJ    L C   A      Def: If J(x) = J(y) →b.
  -CH    L C   A      Def: If H(x) = H(y) →b.
   BHB

A  BHN
```

A OPER L C R B Seg.

```
   D     x            Describe

   NK    x
   NJ    x
   NH    x
   BHN
```

A OPER L C R B Seg.

```
   NH    x            Count variable places

   FEF   L 1   C
A -CPS   1 L   B
  -TU    1     B
   NAH   L
B  FEN   L 1   A
C  BHN
```

Primitives

```
   CPS   x y   b   If E(x) subelement of E(y) →b.
                   (P(x) ⊃ P(y)).
   NAH   x         Add one to H(x)
   TU    x     b   If E(x) is a unit, →b.
```

A OPER L C R B Seg.

```
   NJ    x            Count distinct variables

   AA    1            List for counted-V.
   FEF   L 2   E   F  Find first E of X(x).
A -CPS   2 L   D
  -TU    2     D
   FEF   1 3   C      Find first V of list.
B  CN    2 3   D   CN
   FEN   1 3   B      Find next V of list.
C  SEN   2 1
   NAJ   L         A
   FEN   L 2   A      Find next E of X(x).
E  BHN
```

Primitives

```
   AA    x            Assign an unused list of A(x).
   CN    x y   b      If N(x) = N(y), →b.
```

	CPS	x y	b	If E(x) subelement of E(y), →b. (P(x) ⊃ P(y)).
	NAJ	x		Add one to J(x).
	TU	x	b	If E(x) is a unit, →b.

A	OPER	L	C	R	B	Seg.
	NK	x				Count levels
	TU	L		A		T
	TB	L		B		
	FL	L	1			NK
	NK	1				
	FR	L	2			
	NK	2				
	CKG	2	1	C		CK
	PK	1	L			KL
A	NAK	L				
B	BHN					
C	PK	2	L			KR
	B			A		

Primitives

CKG	x y	b	If K(x) > K(y), →b.
NAK	x		Add one to K(x).
PK	x y		Put K(x) in K(y).
TB	x	b	If E(x) is blank →b.
TU	x	b	If E(x) is a unit →b.

PART 2: Reduction of procedural processes [*S]

The Store instructions that rewrite expressions in various ways can be reduced to processes more like the rest of the primitive set. The new primitives required are (a) two (PA and CP) which belong to types of operations already considered, and (b) four of a new type to manipulate the P sequences. The latter operations insert and delete subsequences from the front end of a given sequence. Thus if P = LRRL and P' = LRRLRLR, then P'' = P' - P = RLR and P''' + P = LRRLRLR. Observe that subtraction can only be performed when the subtrahend is an initial segment of the minuend, and also that addition is not commutative. All these routines involve bringing in the elements, one by one, modifying them and storing them in the new list.

Store a copy of X(x) at (new) A(y) (E(x) = M).

A	OPER	L	C	R	B
	SX	x y			
	AA	C			
	FEF	L	1		B
A	PE	1	2		
	PM	C	2		
	S	2			
	FEN	L	1		A
A	BHN				

Store X(x) in A(y) in place of E(y) (E(y) =V) (take E(x) from w.m.)

A	OPER	L	C	R	B
	SXE	x y			
	FEF	L	1		D
A	CP	L	1		E
	CPS	1	L		C
	PE	1	2		
B	PM	C	2		
	HSPP	L	2		
	HAPP	C	2		
	S	2			

Store X(x) at (new) A(y) as main expression

A	OPER	L	C	R	B
	SXM	x y			
	AA	C			
	FEF	L	1		C
A	CPS	1	L		B
	PE	1	2		
	PM	C	2		
	HSPP	L	2		
	S	2			
B	FEN	L	1		A
C	BHN				

C	FEN	L	1	A
D	BHN			
E	PE	L	2	
	B			B

Store X(x) in A(y) as XL(y).

A	OPER	L	C	R	B
	SXL	x y			
	FEF	L	1		C
A	CPS	1	L		B
	PE	1	2		
	PM	C	2		
	HSPP	L	2		
	HAPL	2			
	HAPP	C	2		
	S	2			
B	FEN	L	1		A
C	BHN				

Store X(x) in A(y) as XR(y).

A	OPER	L	C	R	B
	SXR	x y			
	FEF	L	1		C
	CPS	1	L		B
	PE	1	2		
	PM	C	2		
	HSPP	L	2		
	HAPR	2			
	HAPP	C	2		
	S	2			
	FEN	L	1		
	BHN				

Primitives

AA	x		Assign an unused list to A(x).
CP	x y	b	If P(x) = P(y) →b locates "same" element even though V, G, etc. have been modified).
CPS	x y	b	If E(x) subelement of E(y), →b (P(x) ⊃ P(y)).
HAPL	x		Add a Left to front of P(x).
HAPR	x		Add a Right to front of P(x).
HAPP	x y		Add P(x) to front of P(y).
HSPP	x y		Subtract P(x) from front of P(y).
PA	x y		Put A(x) in A(y).

Conclusion

In this paper we have specified in detail an information processing system that is able to discover, using heuristic methods, proofs for theorems in symbolic logic. We have confined ourselves to description, and have not attempted to generalize in abstract form about complex information processing. Because of the nature of the description, involving considerable rigor and detail, it may be useful to set out in conclusion the main features of LT, especially as these appear to reflect basic characteristics of complex systems.

First of all, LT can be specified at all only because its structure is basically hierarchical, and makes repeated use of both iteration and recursion. So true is this, that one of LT's

main features, the use of a problem-subproblem hierarchy, is hardly visible in the program at all.

LT offers no guarantee of finding a proof; on the other hand, it brings to its task a number of different heuristic methods for achieving its goals. All of these methods are important in making LT sufficiently powerful to find proofs in most cases, and to find them with a reasonable amount of computation, but not all of them are essential. Without chaining, for instance, LT could still function. The methods MSb and MDt still provide it with ways to prove theorems—and even some theorems more easily provable by MCh would yield to the more directly "brute force" approach of the other two.

LT is still a very simple process compared, for instance, with the array of methods, techniques, and concepts used by a human logician. For example, the concepts of commutativity and associativity are nowhere to be found in LT. The analysis of LT and its variations is a subject for later papers. However, the following facts, based on hand simulation, may help put LT in perspective. LT will prove in sequence most of the 60 odd theorems in Chapter 2 of Principia Mathematica. With some extension in the variety of methods and cues employed, it will prove most of the theorems in Chapter 3, in which another connective, "and," is introduced.[16] We know nothing, as yet, about what will be required for an extension to the predicate calculus or to other types of problem solving.

LT uses similarity-testing and matching as a multi-stage search and selection process. The questions of efficiency involved in such processes have already been commented upon in Section II. Additional variation and complexity enters the program through the alternative modes, VV and VCt, for perceiving the logic expressions in the course of testing similarity and of matching.

In these and other ways, the logic theorist is an instructive instance of a complex information process. We expect to learn more about such processes when we have realized the logic theorist in a computer and studied its operations empirically; and when the logic theorist will have been joined by similar systems capable of performing other complex information processing tasks.

References

1. Bowden, B. V. (ed.), Faster Than Thought (London: Pitman, 1953), pp. 181-198.

2. Hilbert, D., and W. Ackermann, Principles of Mathematical Logic (New York: Chelsea, 1950), Chapter 1.

3. Whitehead, A. N., and Bertrand Russell, Principia Mathematica, vol. 1, 2nd ed. (Cambridge: Cambridge U. Press, 1925).

[16] A program to do this has been developed and hand simulated by Mr. Kalman Cohen.

TESTS ON A CELL ASSEMBLY THEORY OF THE ACTION OF THE BRAIN, USING A LARGE DIGITAL COMPUTER

N. Rochester, J.H. Holland, L.H. Haibt, W.L. Duda
IBM Research Laboratory
Poughkeepsie, N.Y.

Abstract

Theories by D.O. Hebb and P.M. Milner on how the brain works were tested by simulating neuron nets on the IBM Type 704 Electronic Calculator. The formation of cell assemblies from an unorganized net of neurons was demonstrated, as well as a plausible mechanism for short-term memory and the phenomena of growth and fractionation of cell assemblies. The cell assemblies do not yet act just as the theory requires, but changes in the theory and the simulation offer promise for further experimentation.

Introduction

The problem of how the brain works can be approached by investigating the elementary components, the neurons, and then seeing how larger and larger assemblies of these operate. Or it can be approached by observing the behavior of the entire organism and working back to determine what the components must be. The former activity is called neurophysiology and the latter is called psychology. Before we can say that the problem is well in hand, these two approaches must meet in the middle so that we have a single consistent picture that firmly connects psychology and neurophysiology.

As the neurophysiologist considers more and more complicated structures of neurons he gets into problems that are less and less related to his normal way of thinking. Curiously, however, some of these problems do not begin to resemble parts of psychology. What is happening is that the neurophysiologist is beginning to think about information handling machines that are too complex to be understood without the specialized knowledge of other disciplines. These other disciplines are information theory, computer theory, and mathematics. People in these other fields need to augment the work of the neurophysiologists and psychologists before the brain can be properly understood.

In the experimental study of the brain it is not yet possible to observe well the electrical interconnections among neurons. No one has yet been able to simultaneously record input and output signals of a single neuron in the brain. For this reason it has not yet been possible to test certain theories about how the brain works by experimentation on animals.

It is possible to measure the electrical characteristics of an isolated neuron in some circumstances.[1,2] One can imagine an elaborate network of such neurons and conjecture on the behavior of the network. The analytical treatment of these networks has proved that one can construct any desired kind of logical machine from elements that are probably much less powerful than neurons.[3,4]

The analytical approach has not been very effective in actually describing the behavior of complicated networks of neurons. However, it has proved effective to simulate such networks and to draw conclusions from the behavior of the simulated network of neurons.

Two sets of simulation experiments were made and another is in progress. In the first of these it was possible to simulate a network of up to 99 neurons and a test was made of part of the theory advanced by D.O. Hebb in his monograph, The Organization of Behavior.[5] The second set tested an unpublished revision of P.M. Milner of part of Hebb's theory with a network of 512 neurons. The third set is to test a further revision. In each case the original neurophysiological theory had to be interpreted in order to get something definite enough to simulate, and these interpretations were done by the present authors.

THE 69-NEURON DISCRETE PULSE SIMULATION

In this paper the term "neuron" will generally be used as an abbreviation for the term "simulated neuron". Likewise the term "synapse" will be used to stand for the term "simulated synapse", in other words for the simulation of the coupling mechanism that enables one neuron to send signals to another. Where ambiguity could arise, qualifying adjectives will be used.

The basic idea of the simulation can be seen by reference to Fig. 1. The large rectangle in Fig. 1 stands for all of the 2048-word high speed electrostatic memory of the Type 701 calculator. The memory was divided into 70 parts, one for each neuron and one for the program. In the area reserved for each simulated neuron were some numbers that might theoretically be measured on a corresponding living neuron. These numbers gave all of the information that was needed about each neuron. Specifically, the things that were known about each neuron, either from its location in memory or from the numbers stored there, were:

1. It's number (name)
2. How long since it had fired
3. How tired it was from having been fired excessively
4. For each of 10 output (efferent) synapses:
   - 4.1 The number of the (efferent) neuron that it simulated
   - 4.2 The magnitude of the signal that it sent to that (efferent) neuron when this (afferent) neuron fired.

Under control of the program, the calculator repeatedly scanned the 69 neurons and, by making calculations, caused these numbers to change as they would have changed if the network had actually been constructed. Therefore, after each pass over the data in memory, the data represented, in great detail, the state of each neuron and synapse in the network at the next instant of time.

In this model, time was quantized into time steps. A neuron could fire at any time step, but not between. A time step corresponded approximately to the interval between the firing of one neuron in a chain to the firing of the next. In the simulation, the average length of time required for a single time step was about 5.3 seconds and this corresponded to perhaps 0.7 milliseconds in the brain. Therefore, the simulation was slower by a factor of 7600.

At any given time step a neuron was either fired or in some state of recovery from being fired. Various recovery curves were used and the one shown in Fig. 2 was typical.

During any given run on the calculator, the neurons were interconnected in a particular net. Each neuron was connected so as to stimulate 10 other neurons. Usually the net was designed by the calculator. It would make a random choice of the neuron to be stimulated by each of the 10 output synapses of each neuron. It would record these choices on punched cards and retain them for the rest of the run.

If a neuron fired at time step (n-1) it would stimulate 10 neurons so as to tend to cause them to fire at time step n. The size of the signals sent to the 10 neurons would depend only upon the fact that the original neuron fired and upon the magnitudes of the interconnecting synapses. To say this another way, the input signals to a neuron, together with its threshold, would determine whether or not the neuron would fire, but if it did fire, the strength of firing would not depend on the input signals.

The input situation of a typical neuron is shown in Fig. 3 with some possible values of synapse magnitudes. The behavior of neuron x is shown in the following table.

Neurons that fired on step (n-1)	Mag	Threshold of x	Would x fire on step n ?
AB	295	256	Yes
BCE	252	256	No
BCDE	336	256	Yes
ABCDEFG	839	839	Yes
ABCDEFG	839	938	No
ACF	408	376	Yes
EFG	376	376	Yes

It can be seen that the input circuits to such a neuron can provide quite sophisticated switching.

Not all of the properties of the simulated neurons have been described. However, to make the exposition easier to follow, it is convenient to skip ahead and show some observations on the behavior of networks of neurons. Except for some minor difficulties, this behavior would be obtained with neurons like those already described. The discussion of these minor difficulties will be clearer after showing these results.

Fig. 4 shows an example of what will be called diffuse reverberation. Each row in this figure indicates with a 1 those neurons that fired and with an 0 those neurons that did not fire in a particular time step. Each column, of the 64 columns at the right, shows the history of a single neuron. The right hand 64 columns of Fig. 4 show, therefore, the complete firing history of 64 neurons for 50 time steps.

Fig. 5 shows, as a function of time, the number of neurons that were simultaneously fired. The time covered here is a little larger

than in Fig. 4 and shows the complete history beginning with a quiescent net and continuing until the activity died out.

We propose this diffuse reverberation as a plausible mechanism for short term memory, the kind of memory that is involved in remembering the intermediate results in mental arithmetic. We will discuss later some conjectures as to how the brain can make use of such a memory mechanism.

Now another property of the neurons will be described. When neuron A participated in firing neuron B, the synapse that enabled A to stimulate B was increased in magnitude unless it already had reached the limit of 938, in which case it remained constant. This characteristic was our version of Hebb's basic neurophysiological postulate. Hebb postulated that, "When an axon of cell A is near enough to excite cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."[6]

This property of simulated neurons is somewhat curious. No process of just this sort has been observed in living tissue. However, it has not been possible to demonstrate, by measurement, that the Hebb postulate is false. Nothing else has been observed that could account for learning and memory in a plausible way. The Hebb postulate suggests a plausible machine that does not contradict experiment.

The purpose of the assumption about the growth of synapses is to get a mechanism for the retention of long term memory. When an animal experiences some event there will be activity in its brain. This activity will consist of a spacial and temporal pattern of firing of neurons. During the experience, the synapses involved will be strengthened, according to Hebb's postulate. Therefore, the same, or a similar, sequence of neural events is more likely to take place later than it would have been if the animal had not had that experience. A repetition of some part of the neural events that were associated with an experience is assumed to be the act of recalling the experience. It is evident that the mechanism that Hebb postulated would tend to cause recollections. The question of whether or not the postulate is sufficient is, in a sense, the main topic of this paper.

If no additional rule were made, the Hebb postulate would cause synapse values to rise without bound. Therefore, an additional rule was established: The sum of the synapse values should remain constant. This meant that, if a synapse was used by one neuron to help cause another to fire, the synapse would grow. On the other hand, if a synapse was not used effectively, it would degenerate and become even less effective, because active synapses would grow and then, to obey the rule about a constant sum of magnitudes, all synapses would be reduced slightly, so the inactive synapses would decrease.

Before discussing network action further, another property of the neurons will be mentioned. A neuron fired at too high a frequency becomes less sensitive, so that more stimulation is required to fire it. The effect of this is shown in Fig. 6, which shows the threshold as a function of time when the neuron is fired repeatedly with a constant level of stimulation. As with a living neuron, this simulated neuron fires rapidly at first and then settles down to a lower rate of firing.

This process is called fatigue because of the obvious analogy to living neurons. A significant aspect of fatigue is that it is a form of memory and, as such, may plan an important part in the operation of the brain

The concept of cell assembly occupies a key position in Hebb's theory. A cell assembly is a group of neurons that are interconnected in a very complex fashion and within which diffuse reverberation can take place. Fig. 4 shows just such a situation.

Parts of the cortex are imagined to consist of a large number of cell assemblies, each of which contains a large number of neurons. Only a small fraction of the cell assemblies are aroused at any one time. In other words signals are reverberating in only a few cell assemblies at once. Just which cell assemblies would be aroused at any one time would depend in large part upon what cell assemblies had been aroused at a previous instant of time, and in small part upon signals from elsewhere.

In the language of information theory, this part of the brain can be considered to be a finite state transducer, in which the internal state is determined by noting which cell assemblies are aroused and which are quiescent. In other words, the brain should exhibit a kaleidoscopic sequence of patterns of cell assembly arousal. It is outside the scope of this paper to expound Hebb's theory, so it will be assumed henceforth that the reader either understands the significance of a finite state transducer or has read Hebb's book.

In passing, it is worthwhile to point out how appropriate the finite state transducer description is for Craik's "hypothesis on the nature of thought."[8]

Hebb's theory required that it be possible for a neuron to belong to several different cell assemblies and that not all of these assemblies be aroused at once. Hebb's theory also required that it be possible for a neuron to change its affiliation from one cell assembly to another. It may be possible to devise a theory that has only the second requirement, but no further consideration of this possibility will take place in this paper.

The problem of how cell assemblies can arise and how they become modified, is vital to this theory. It will be shown that Hebb's scheme is unlikely to work with neurons of the type described so far. It will also be shown that, by suitably improving the neurons and by making the network more complex, cell assemblies can be made to form spontaneously. It will further be shown that these cell assemblies are not entirely satisfactory but that there is a plausible course for further investigation.

Suppose that there is initially some activity in a network of neurons and that input signals are impinging on the network. Suppose also that from time to time a particular input signal, S, arrives. When S first arrives, it will impinge on some internal state, I_j. In other words it will impinge upon some particular configuration of states of individual neurons. The particular sequence of internal states, I_{j+1}, I_{j+2}, I_{j+3}, ..., that follows will strengthen certain synapses in such a way that the sequence I_j, I_{j+1}, I_{j+2}, I_{j+3}, ... is more likely to occur again. It was conjectured that the next time S occurred, some part of I_j would be in existence and that some part of the sequence I_j, I_{j+1}, I_{j+2}, I_{j+3}, ... would be reinforced. As S appeared repeatedly some characteristic response to S gradually would become sufficiently reinforced as to be identifiable. As the characteristic sequence was arising there would appear, in it, points where diffuse reverberation could occur. In other words there would be some internal state I_{j+k} which would repeat some part of an earlier state in the sequence. As soon as this happened the rate of reinforcement of the connections would increase, because each time the stimulus S occurred the sequence of states would be such as to give several reinforcements to some of the connections instead of just one reinforcement. It was conjectured that cell assemblies corresponding to some common stimuli would arise in the brain in this way.

In order to test this conjecture about the manner in which cell assemblies form, a program was written to generate an appropriate environment for the neuron network, and an arrangement was set up for the network to receive signals from the environment. To receive the signals, six neurons were chosen to act as receptors. It was arranged that no neurons would stimulate these receptors. Instead, they could be fired only by an external program to enable the calculator to reach in and modify the one bit on each of the six neurons that indicated whether or not it had just received enough stimulation to fire. The synapses from the receptors spread out diffusely through the network.

The neuron net was stimulated once every ten time steps with a 6-bit signal that could define the state of each of the six receptors. The signals were chosen by a program whose action is illustrated in Fig. 7. It is a Markov process in which there is some probability that the input will be random but mixed in with the random signals are frequent occurrences of certain sequences. The network then had the opportunity to develop a characteristic response to each of the three sequences.

The network did not develop any characteristic responses and there was no sign of development of cell assemblies. A number of variations on this experiment were tried, all with the same result. Then the reason for the difficulty was realized and a simulation experiment was run to verify the explanation.

In such a neuron network, the idea that a detailed temporal-spacial pattern of firing can be effectively reinforced by a partial repetition is false. The reason can be seen from the following experiment. A simulation experiment was run to a convenient point where diffuse reverberation was taking place. Then all the data was punched on tabulating cards. These cards contained all relevant information so that, if they were read by the calculator, the simulation would go on from where it left off. Before the cards were read by the calculator, however, they were reproduced to give four identical decks of cards. Then three of these decks were slightly modified, each in a different way. In each case the modification was to choose some neuron that was about to be fired and manually change the number that specified its state of recovery so that it wouldn't fire quite so soon. Then four simulation experiments were run, one with each deck.

The four sets of results were compared and it was found that the detailed patterns of firing diverged rapidly. In just ten time steps, in each case, over 30 per cent of the neurons firing were different. This result is shown in Fig. 8. This shows that even slight differences rapidly grow to be large differences so there is little chance that a detailed pattern of firing can be effectively reinforced.

It was concluded from this work that some additional structure was needed within a network to allow all assemblies to form. A plausible model of a short term memory had been demonstrated but rather convincing evidence had been found to show that Hebb's postulate was not enough to make cell assemblies form.

Some other experiments were run which coincided in time with the work of Farley and Clark[9] and which reached essentially identical results. However, these did not seem to throw any light on the central problem of how the brain works, so this line of investigation was dropped.

512-Neuron F.M. Simulation

At this point we conferred with D.O. Hebb and one of his people, P.M. Milner. Milner had been working on a revision of part of Hebb's theory to introduce more recent neurophysiological data. The essence of Milner's idea was that inhibitory synapses, as well as excitatory synapses, are needed and that within a cell assembly most synapses are excitatory, while between cell assemblies most synapses are inhibitory. This idea sounded to us like a plausible cure for the troubles in the first model. It made engineering sense.

The significance of the idea can be seen by considering two cell assemblies. These will act like an Eccles-Jordan Flip Flop circuit. Suppose one is aroused. It keeps itself going by its internal excitatory connections and keeps the other quiescent by the inhibitory interconnections. Finally it begins to fatigue. As it begins to falter, it inhibits the other less strongly, so sporadic residual activity in the other begins to increase. This in turn inhibits the aroused cell assembly, causing it to falter more. This feedback condition causes an abrupt switching so that the aroused one becomes quiescent and the quiescent one becomes aroused. A more detailed discussion of this can be found in Appendix 1.

It seemed certain that the switching action would take place, but it was not clear whether the possibility of having inhibitory synapses would be enough to allow cell assemblies to arise or whether some cell assembly structure would have to be built in at the start.

Experiments with the discrete pulse model indicated that diffuse reverberation was a fairly reliable sort of thing in a net of 63 neurons, but quite erratic in a net with 21 neurons. Therefore it was felt that in a new experiment there should be a larger number of neurons in a net. A major obstacle to this was that the calculator was not fast enough to manage a very much larger net, even though this was to be done on the Type 704 which is faster than the 701. Something had to be sacrificed.

It was decided to sacrifice the knowledge of exactly when an individual neuron fired. All that the machine or the experimenter could know was the frequency at which a neuron was firing, and not the exact instants of time at which it did fire. The frequency would vary from time to time, so this was called the FM model.

One particular version of the FM model will be described here. There were 512 neurons, each with 6 input (afferent) synapses and a number of output (efferent) synapses that varied from one neuron to another. The synapse magnitude lay between -1 and +1 and changed as long term learning took place. The frequency of a neuron varied from 0 to 15. Equations are given in Appendix 2 to specify precisely how these quantities varied from time to time, and a qualitative description is given below in the text.

The magnitude of a synapse was much like a correlation coefficient between the two neurons that it connected. If the frequencies of the two neurons usually went up and down together, the synapse magnitude would grow toward +1. If, on the other hand, one neuron was usually inactive while the other was active, the synapse magnitude would approach -1. This is the FM version of Hebb's basic neurophysiological postulate.[6]

The frequency of a neuron was obtained essentially by calculating, for each synapse, the product of the synapse magnitude and the frequency of the stimulating (afferent) neuron, adding these products, and normalizing. It was further bounded by not being allowed to go negative. Therefore a neuron could have a high frequency only if it was stimulated through positive synapses by neurons with large frequencies and not simultaneously stimulated through negative synapses by neurons with large frequencies .

The fatigue increased if the frequency was high; stayed constant if the frequency was intermediate; and decreased if the frequency was low. Furthermore, it was not allowed beyond the bounds of 0 and 7. A fatigue of 7 could nearly stop a neuron while a fatigue of 3 did little to it.

An important change was made in the nature of the connections in the net. A distance bias was introduced so that two nearby neurons were more likely than two remote neurons to be connected together through a synapse. In the experiment described in this paper, the neurons were visualized as being arranged in a cylinder, as shown in Fig. 9. The cylinder was 16 neurons high and 32 neurons around. If two neurons were within eight of each other, they were as likely to be connected by a synapse as any other two neurons that were within eight of each other. However, no neurons that were farther apart were connected by synapses.

Four blocks of four neurons each were selected to act as receptors. These four blocks are shown in Fig. 9. The procedure that was used most of the time was that receptor areas 1 and 4 were controlled to have maximum activity for three successive time steps and then the net was allowed to operate with no external stimulation for three time steps. Then areas 2 and 3 were controlled to have maximum activity for three time steps and then the net was again left alone for three time steps. This cycle was repeated many times. The cycle was considered to be the equivalent of about 0.2 seconds in an animal and took about 160 seconds on the calculator.

Cell assemblies did actually build up around each of the receptor areas. Within a cell assembly the interconnections were largely excitatory and between cell assemblies they were largely inhibitory.

The activity of each neuron at each time step for one complete cycle is given in Table 1.

Table 1

Activity During One Complete Cycle of Stimulation

1.

```
004000000000060000001403141003 00
004000031001000000770104000000 00
000000207000000000770100000120 00
000000000000500100000500030000 00
001000530110040000061040002000 00
000001010004000000050011000000 00
000000000100004000100001100000 0
000000100000000000011000100000 00
000004000030000300000000000000 01
000000010000004105000000004001 10
000000000000006000000041000000 00
000000000000060000000000000000 00
000003300000001000000001010000 4
000000000500101003770000000100 00
000000000000000000770000000010 00
000000005010000010000001000001 0
```

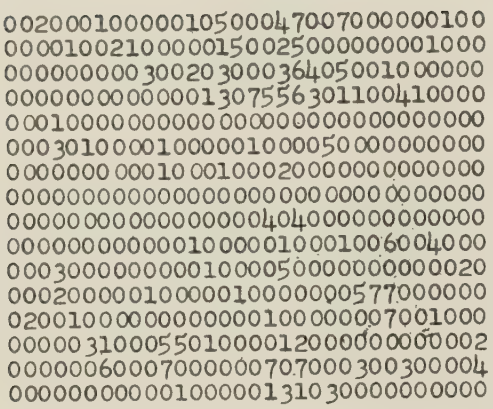
2.

```
004002000000050001600000100003 00
003020042002000200770603000100 00
000000050000020007750700000000 0
000000030000020005000000100200
000004100000000000000000100000
000001031000000000020201000000 00
000400000010001000000000010000 00
000000101000000000000000000000 00
000000000010000100070000000000
000000011000001002000004000000 00
000000000100004100410000000000
000000000000031001300000200000 00
000002300000000000000000000000 00
000000100003100004770000000000 00
000102000000000004770000002000 00
000001000000000000000000000000 00
```

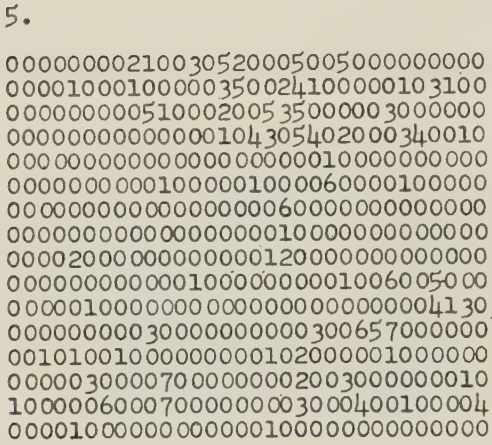
3.

```
003001110000012006700700000010 0
002020033100000200770100200000 00
000000000000103000775070000000 00
000000012000002053730001013000 00
000000000000000000000000000000 00
000300030010000010000301000000 00
000300000000001000000000000000 00
000000100000000000000000000000 00
000000000000000013070000000000 0
000000010000100100100011060000 00
000300000000200004000001000000 0
000000000000000001100000670000 00
000000200000000000000001070000 00
000001100005000011770000000000 0
000001300050010007770000003000 3
000001000000000001300300000000 0
```

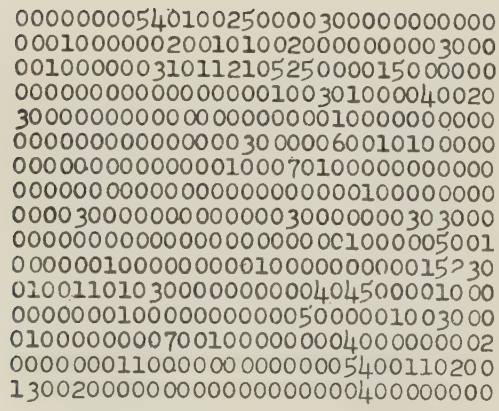
4.



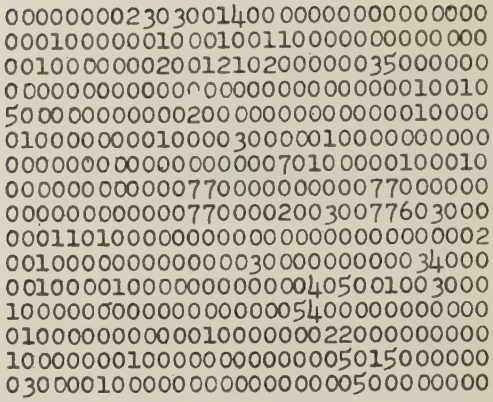
5.



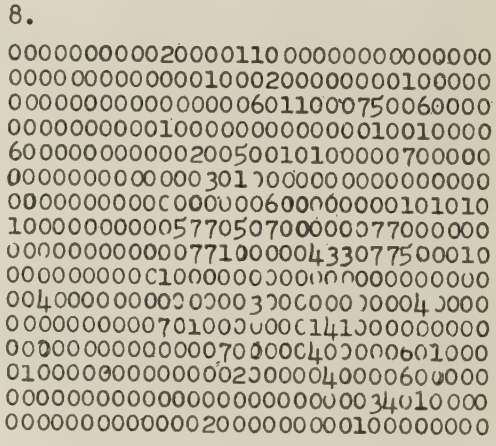
6.



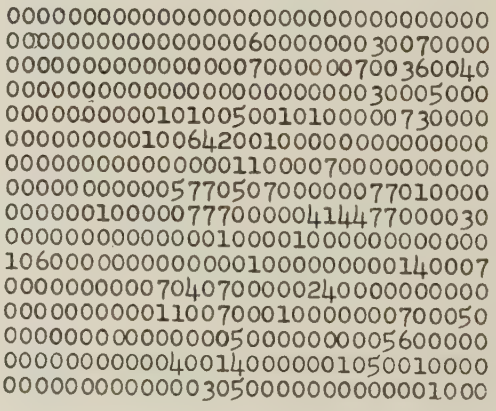
7.



8.



9.



10.

```
0000000000000000100000000001006
0000000000000000600000600070000
0000000000000007000000710450040
0000000000000005000000000005000
0003000000300036020000040750000
0000000010006530020000000000000
4000000000000031000070020000000
0000000000067706070001020000000
0000060000003700000413170050030
0000000000000040000600000000100
1000000000000001000700000000007
0000000010535070000000000000000
0000000001330500060000040600061
0000000000000006000020005501300
0000100000040047000020003002000
0000000003300405000000000013000
```

11.

```
0000000000000000410002040010300 5
0010000000000000700000602060000
0000002000000000200002020051005 0
0010000000002006000000000005000
0000000300500442040230003006000 0
0000000130371400013003000000000
5000000000000053002070030020000
0000000000011601020010004000100 0
0000070000003400000102200050000
0000000000000040000600000044000 0
2000000000010020120170500110000 5
0000000010032060000000000010100
0000010003360000050001070000140
0000000070000040000020004014300
0000000000040056000000020013000
0000000007400206000000100003000 0
```

12.

```
0000000000000100640004060300500 6
0010000000000000100005403000000
0000000040000000100004000011001 0
0010000000003006000004000000000
0000004500400540040440305016000 0
0000000000051010010400500001000 0
4000000010100004400000003005010 0
0000000200022010000200130000000
0000004000300001000000200030402
0000000000000020000310020074000 0
0000000000000401402606300100000
0000000000030400000000000130032
0000031001312000003000304014000 4
0000000007000000000030100013300
0000000000000031100010000010000
0000000052103000000010000300 01
```

In Table 1 the numbers are just half of the frequencies of the neurons. This was done to reduce the cost of printing. Since it is difficult to see what is going on without practice and effort, a small section of the sixth and ninth steps have been reproduced side by side in Fig. 10. These times were chosen to contrast the arousal of 1 and 4 which suppress 2 and 3, with the arousal of 2 and 3 which suppress 1 and 4. Nearly every neuron has chosen allegiance to one cell assembly or another. Only 3 of the 224 neurons shown are active at both times.

Examination of the synapses also showed that cell assemblies had formed. A dividing line was found between area 1 and area 2. Synapses that crossed this line were predominately negative while synapses that failed to cross it were predominantly positive.

There is no doubt that cell assemblies did form. A very detailed statistical study of the allegiance of neurons to cell assemblies was not made because, as will soon be evident, the model still needs improvements, and the statistics of the improved model would be different.

A further significant characteristic needed by the Hebb-Milner theory was evident. Over sequences of 100 or so time steps (perhaps the equivalent of 17 minutes) neurons were observed to change allegiance from one cell assembly to another. In other words, "fractionation" and "recruiting" were observed.

Some evidence was found to indicate that one cell assembly tended to arouse another. However, the tendency was weak. It can be seen that the only possible excitatory synapses between cell assemblies would be those involved with neurons of dubious allegiance. Apparently this was insufficient to allow spontaneous activity in the net. The only activity was caused by the input signals and the arousal of cell assemblies was completely determined by the input. The theory requires that the preceding central activity (set) be much more influential than the input stimuli, so clearly some changes are needed.

Plans for the Future

After studying the detailed results of this experiment, we arrived at a conjecture as to what should be done next. This conjecture was based on our intuition gained from experience in designing computing machines. We felt that the inhibitory synapses should be separated from the excitatory synapses and should follow different rules. Appendix 1 describes some detailed con-

siderations of the transmission of activity from cell assembly to cell assembly.

We then consulted again with P. M. Milner and learned that he had just produced a further revision of the theory that had just this property of synapses with differeing characteristics. His new model appears also to have the characteristic that the cell assemblies would be much more diffuse than in the FM Model described here. This would correspond better to what is expected in the brain and would make a better machine because one cell assembly could directly affect a larger number of others. It is not within the scope of this paper to discuss this new scheme because we have not yet reduced it to our terminology and tested it. However, the work is proceeding.

Summary

The first set of experiments, designed to test parts of the theory advanced in The Organization of Behavior, by D. O. Hebb, simulated a network of 69 neurons with a "Discrete Pulse Model." This set of experiments clearly illustrated the diffuse reverberation that is advanced as an explanation of short term memory. There was, however, no tendency for neurons to group into cell assemblies.

The second set of experiments were designed to test P. M. Milner's revision of Hebb's theory with an "F. M. Model" which kept track of the frequency of firing of 512 neurons but ignored the precise timing of individual firings. Cell assemblies formed and exhibited the "fractionation" and "recruiting" required by the theory. The cell assemblies, however, were not able to arouse one another, so this model was too heavily dominated by environment.

A third set of experiments is in progress. It is hoped that this set will get around the next major obstacle in producing a model that will do what the neurophysiological theory requires.

This kind of investigation cannot prove how the brain works. It can, however, show that some models are unworkable and provide clues as to how to revise the models to make them work. Brain theory has progressed to the point where it is not an elementary problem to determine whether a model is workable. Then, when a workable model has been achieved, it may be that a definitive experiment can be devised to test whether or not the workable model corresponds to a detail of the brain.

Appendix 1.

The Interaction of Cell Assemblies

Suppose that all synapses within a cell assembly are excitatory and that both excitatory and inhibitory synapses go between all assemblies. Suppose also that the effect of stimulation at a synapse rises suddenly when the preceding (afferent) neuron fires, and then dies out more slowly. For example, a model of this could be a chemical transmitter that was discharged on the stimulated (efferent) neuron and that was destroyed at an exponential rate. Suppose also that the effect of an excitatory synapse fades more slowly than the effect of an inhibitory synapse. In terms of the chemical transmitter, this could mean that two different chemicals were used for inhibition and excitation, and that these were destroyed at different rates. Finally, suppose that the total inhibitory stimulation of an aroused cell assembly on a quiescent cell assembly dominates the total excitatory stimulation.

While a cell assembly is firing actively it will suppress its neighbors. However, when its neurons tire and it begins to falter, the inhibition will drop more rapidly than the excitation. When the level of inhibition drops below the level of excitation, switching will take place.

This sort of interaction between neurons is being built into the third set of experiments.

Appendix 2
Equations Describing the FM Model

The structure of the net is given by

$$j = g(h, i)$$

where i is the number of the efferent neuron, j is the number of the afferent neuron, and h is the number of the afferent synapse for the ith neuron. g (h, i) is determined at the beginning of an experiment, and remains constant.

The following quantities for all i, j determine the state of the model at any time t.

Symbol	Number of Bits	Description
x (i, t)	4	frequency of neuron i at time t
x (i, t)	4	average frequency of neuron i at time t
d (i, t)	3	fatigue of neuron i at time t
r (i, j, t)	8	magnitude of the synapse at time t coupling stimulation from neuron i to neuron j
R (i, t)	8	a function of x(i, t), x(i, t-1), ...

Initial conditions for the net are given by the values $x(i,0)$, $\bar{x}(i,0)$, $d(i,0)$, $r(i,j,0)$, and $R(i,0)$.

The quantities $S(i,j,t)$ and $x'(i,t)$ are intermediate results in the calculation. A single time step consists of the successive evaluation of the following formulas:

1) $S(i,j,t) = r(i,j,t)\sqrt{R(i,t)\,R(j,t)}$

2) $R(i,t+1) = (1 - \frac{1}{m})\,R(i,t) + (x(i,t) - \bar{x}(i,t))^2$

$R(j,t+1) = (1 - \frac{1}{m})\,R(j,t) + (x(j,t) - \bar{x}(j,t))^2$

$m = 32$

3) $S(i,j,t+1) = (1 - \frac{1}{m})\,S(i,j,t)$

$+ (x(i,t) - \bar{x}(i,t)) \cdot (x(j,t) - \bar{x}(j,t))$

$m = 32$

4) $r(i,j,t+1) = S(i,j,t+1) \Big/ \sqrt{R(i,t+1)\;R(j,t+1)}$

5) We define $p(i,t) = j$ such that $r(i,j,t) \geq 0$, and $q(i,t) = j$ such that $r(i,j,t) < 0$, including only values of j such that the synapse (i,j) exists.

Then

$x'(i,t+1) =$

$$k_o \left[\frac{\sum_{p(i,t+1)} \left[r(i,j,t+1) + k_1 \right] x(j,t)}{\sum_{p(i,t+1)} r(i,j,t+1) + k_1} - \frac{\sum_{q(i,t+1)} \left[r(i,j,t+1) - k_1 \right] x(j,t)}{\sum_{q(i,t+1)} r(i,j,t+1) - k_1} \right]$$

$k_o = 1.25$, $k_1 = 1/512$

6) $d(i,t+1) = f(d(i,t), x'(i,t+1))$

$d(i,t)$ \ $x'(i,t+1)$	0-4	5-9	10-15
0	0	0	2
1	0	1	3
2	1	2	4
3	2	3	5
4	3	4	6
5	4	5	7
6	5	6	7
7	6	7	7

7)

$$x(i,t+1) = \begin{cases} \text{an externally controlled value, when the neuron } i \text{ is a stimulated receptor} \\ X(x'(i,t+1), d(i,t+1)) \text{ when neuron } i \text{ is not a stimulated receptor} \end{cases}$$

Table of $x(i,t) = X(x'(i,t), d(i,t))$

x' \ d	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	0
2	2	2	2	2	1	1	1	0
3	3	3	3	2	1	1	1	0
4	4	4	4	3	3	2	1	1
5	5	5	4	4	4	3	2	1
6	7	6	6	4	4	3	2	1
7	7	7	7	6	5	4	2	1
8	8	8	8	7	5	5	2	1
9	9	9	9	7	5	5	2	1
10	10	10	9	8	6	5	3	1
11	11	11	11	8	6	5	3	1
12	12	12	12	9	6	6	3	2
13	13	13	12	9	7	6	3	2
14	14	13	13	10	7	6	4	2
15	15	15	13	10	7	6	4	2

8) $\bar{x}(i,t+1) = (1 - \frac{1}{m})\,\bar{x}(i,t) + \frac{x}{m}(i,t+1)$

References

1. Brink, F. Jr. "Excitation and Conduction in the Neuron" and "Synaptic Mechanisms". pp. 50-120 in Handbook of Experimental Psychology, Ed. by S. S. Stevens, John Wiley and Sons, Inc., New York; 1951.

2. Eccles, J. C., The Neurophysiological Basis of the Mind, Oxford: The Clarenden Press, 1953.

3. McCulloch, W.S., and Pitts, W., "A Logical Calculus of the Ideas Immanent in Nervous Activities", Bull. Math. Biophysics, vol. 5, pp. 115-133, 1943.

4. Kleene, S. C., "Representation of Events in Nerve Nets in Finite Automata", in Automata Studies, Annals of Mathematics Studies, No. 34. Ed. by C. E. Shannon and J. McCarthy. Princeton: Princeton University Press, 1956.

5. Hebb, D.O., The Organization of Behavior, New York: John Wiley and Sons, Inc., 1949.

6. Ref. (5), p. 62.

7. Ref. 2, and notice that Hebb's postulate (Ref. 6) is not necessarily related closely to Eccles "post-tetanic potentiation". On p. 196 Eccles shows the effect of a million volleys (Fig. 6A, 36 minute curve) and this is much more severe than is relevant for the present discussion.

8. Craik, K. J. W., The Nature of Explanation, Cambridge: The University Press, 1952.

9. Farley, D. G., and Clark, W. A., Proceedings of the Western Joint Computer Conference, 1955.

0	1	2	3	4	5	6	7
8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23
24	25	26	27	28	29	30	31
32	33	34	35	36	37	38	39
40	41	42	43	44	45	46	47
43	49	50	51	52	53	54	55
56	57	58	59	60	61	62	63
64	65	66	67	68			
PROGRAM							

Fig. 1 - Allocation of memory.

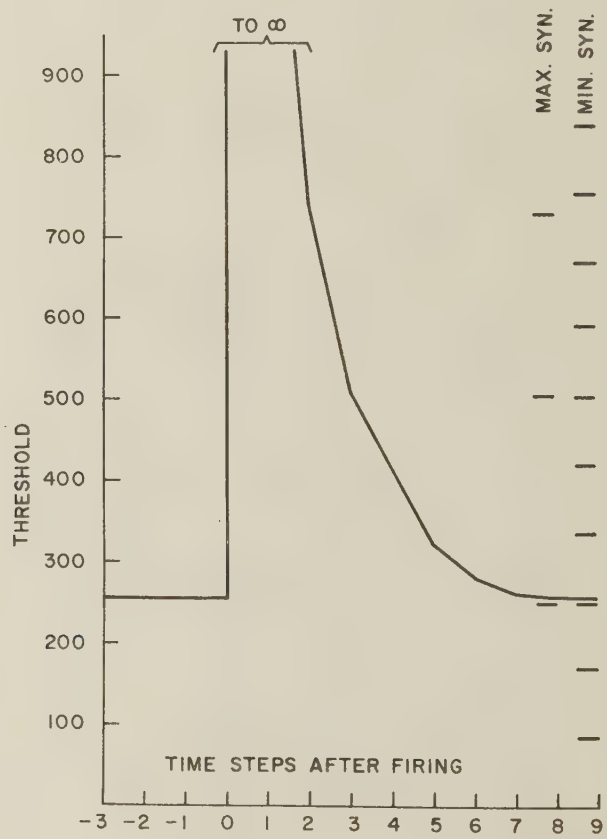


Fig. 2 - Threshold curve.

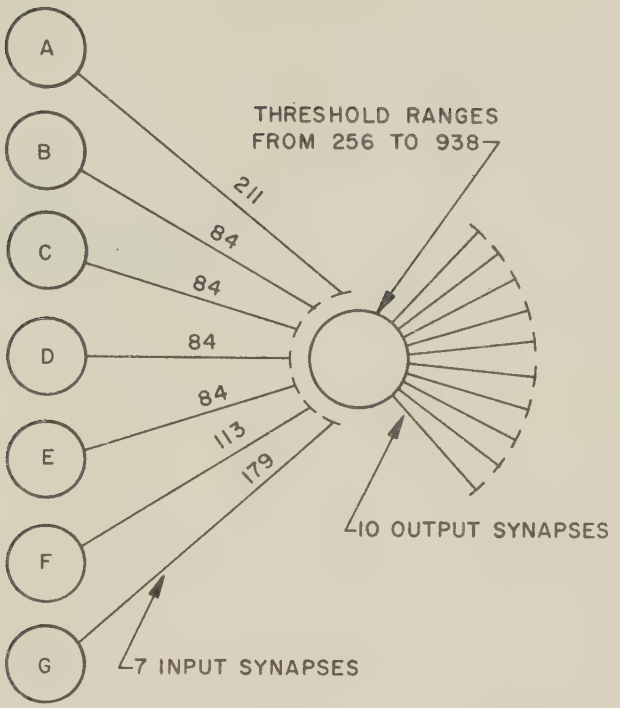


Fig. 3 - Example of a simulated neuron and its connections.

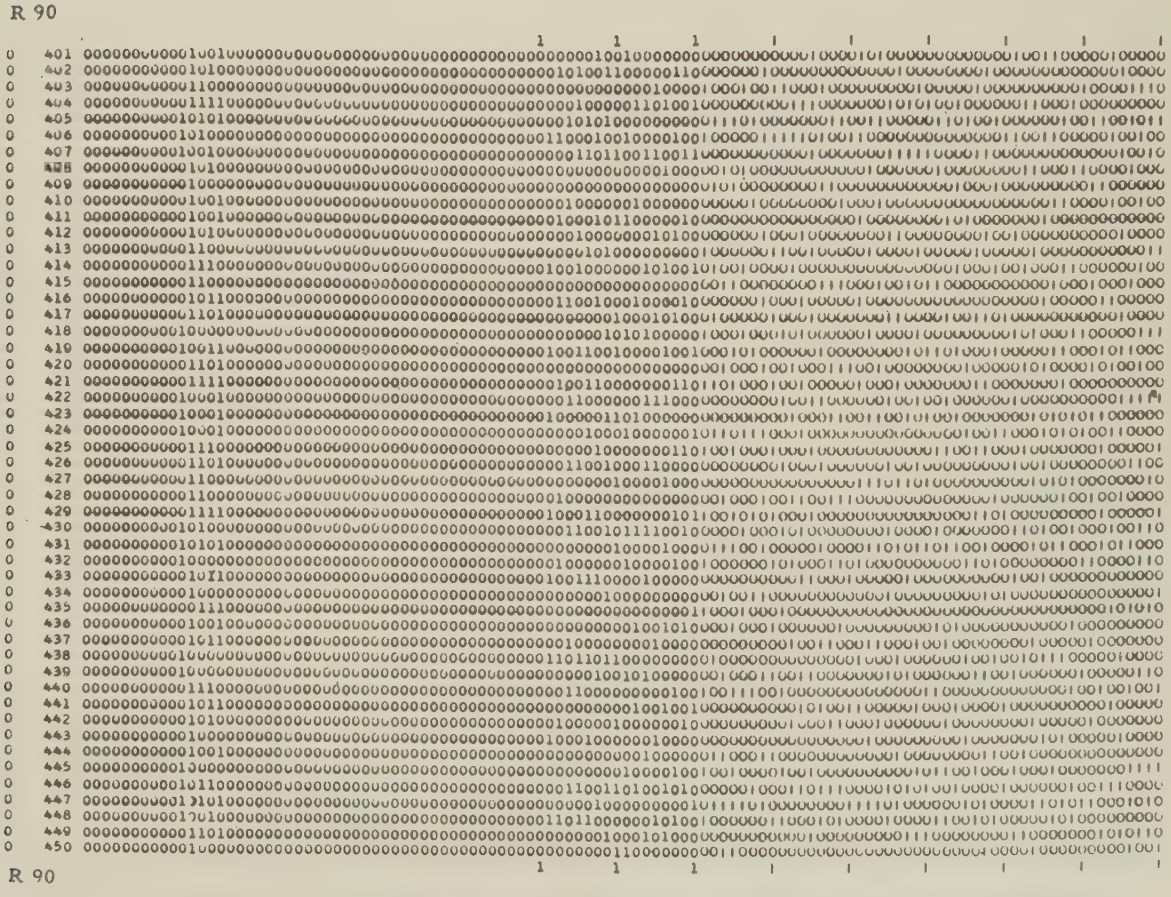


Fig. 4 - Firing pattern of 64 neuron for 50 time steps showing diffuse reverberation.

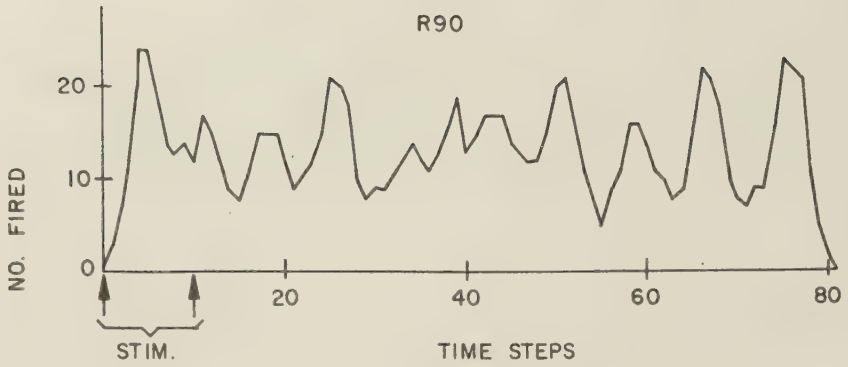


Fig. 5 - Number of neurons firing at each time step showing diffuse reverberation.

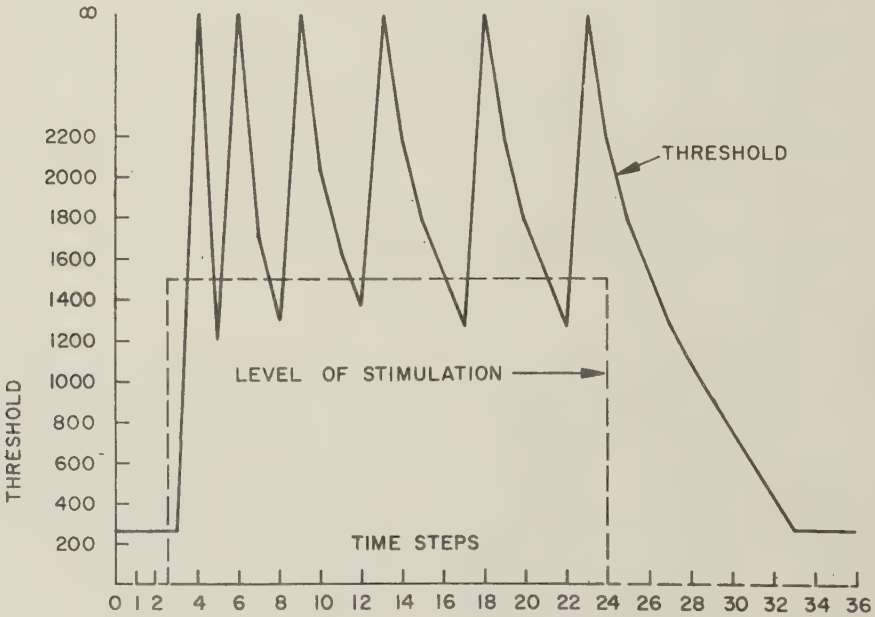


Fig. 6 - Threshold as a function of time.

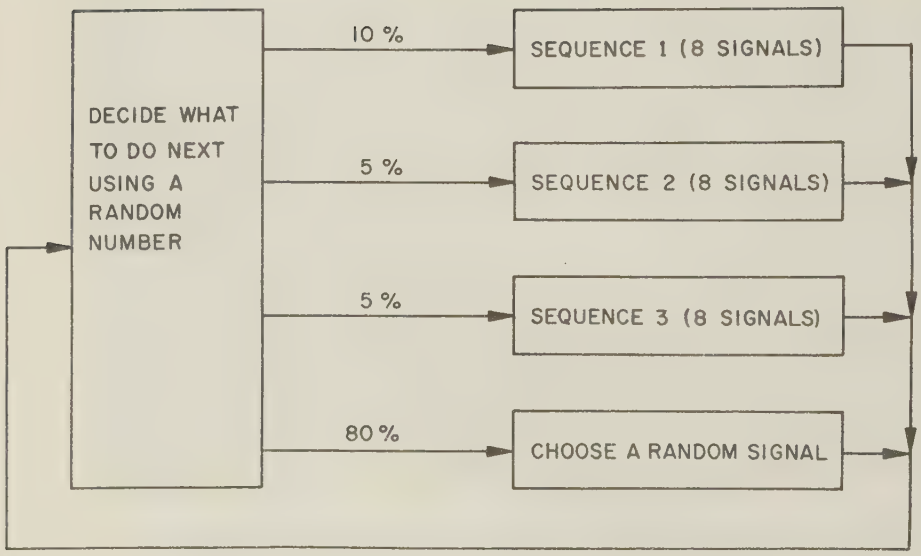


Fig. 7 - Environment.

Time Step	Total Neurons Firing C Cntl. Run	A N40 Sup.	B N61 Sup.	D N70 Sup.	Different Neurons Firing A•C	B•C	D•C
151	31	30	30	30	1	1	1
152	30	30	28	28	0	2	2
153	28	29	27	27	1	3	5
154	30	31	29	31	1	9	5
155	31	31	28	30	2	19	9
156	33	34	26	30	3	27	17
157	30	29	31	31	5	37	19
158	31	26	32	29	9	37	16
159	32	30	32	34	14	32	22
160	34	32	33	38	20	35	22

Three separate runs are represented in this chart in addition to the control run, C. In run A, neuron 40 was suppressed; in B, N61 was suppressed, and in D, N70 was suppressed.

Fig. 8 - Divergence after suppressing one firing of one neuron.

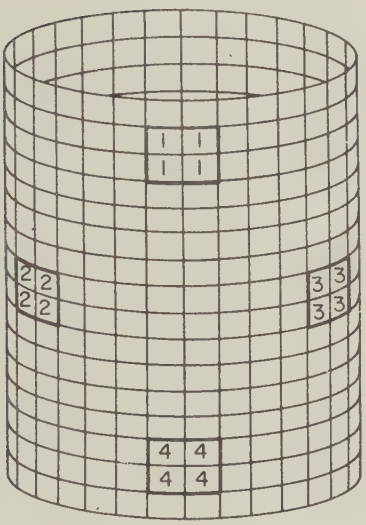


Fig. 9 - Arrangement of neuron in fm model.

00	00	20	50	00	00	00	00	30	00	00	00	00	00
00	10	00	10	06	00	20	00	00	00	00	00	03	00
10	10	20	10	07	50	20	50	00	00	00	07	10	50
00	00	00	00	00	10	00	00	30	00	10	00	03	00
01	00	00	05	00	00	01	00	01	10	00	00	00	00
06	04	02	30	00	01	00	00	00	60	00	00	10	00
00	00	10	01	01	00	70	00	10	07	00	00	00	00
07	07	00	05	00	07	00	00	00	00	00	10	07	07
07	07	07	00	00	00	30	00	04	01	04	04	07	07
00	00	01	00	00	00	00	01	00	00	10	00	00	00
00	00	00	00	11	00	00	00	00	00	00	00	00	00
04	00	07	00	00	00	00	00	42	04	40	50	00	00
00	00	07	00	00	00	01	50	00	00	00	00	00	10
00	10	00	05	00	00	00	00	00	00	40	00	00	05
00	00	01	04	00	00	00	00	00	00	51	40	05	00
00	03	00	05	00	00	00	00	00	00	00	40	00	00

Fig. 10 - Illustration of cell assemblies.

THE MEASUREMENT OF THIRD ORDER PROBABILITY DISTRIBUTIONS OF TELEVISION SIGNALS

W. F. Schreiber
Research Department, Technicolor Corporation
Burbank, California

Summary

A device has been built for the rapid, automatic measurement of the third order probability density of video signals. Special cathode ray tubes are used to perform a 64-level amplitude analysis on each signal. A triple coincidence is formed among the three analyzer outputs, the number of occurrences in a frame interval being stored in an 8 Mc counter. These numbers are recorded on magnetic tape which then becomes the input to an electronic computer. The computer calculates the conditional entropy, i.e., the information generated by one of the signals when the other two are known. Examples are presented of second and third order distributions, and of entropies calculated for a variety of scenes.

Introduction

One standard television signal occupies four times the bandwidth allotted to the entire A-M broadcast service. This is undesirable for many reasons. In the first place spectrum space is in short supply with many communication facilities competing for frequency assignments. Secondly there are some applications of television which are made considerably more difficult on account of the wide bandwidth, for example, recording. Finally, there are other television applications which are made quite impossible, for example, long distance wireless transmission for both military and civilian use. It seems worthwhile to investigate methods by which the bandwidth might be reduced.

There are two categories of techniques which might be called upon to reduce the bandwidth of television transmissions. The first depends upon the psychology of vision. Such techniques involve removing from the television signal certain information not required for a satisfactory image to be reproduced, and to do this in some way which results in a bandwidth compression. Vertical interlace as used in commercial television transmissions is an example of this kind of bandwidth reduction. It results in a full definition picture with half the bandwidth otherwise necessary. Other schemes of this type have also been proposed, such as Toulon's suggestion for "knight's move" scanning.[1] The second category of bandwidth reduction techniques depends upon information theory; that is, these techniques attempt to exploit statistics of the television signal so as to remove redundancy and to code the signal in such a way that the transmission bandwidth is reduced while at the same time the picture presented to the observer is essentially unchanged. Again there have been many suggested schemes but as yet none has been instrumented. Now just as techniques based on the psychology of vision must depend on carefully made psychophysical experiments, methods depending on the statistics of the signal must have as their basis a detailed knowledge of the statistics of television signals. This paper describes a machine for rapidly and accurately measuring those statistical parameters of television signals which are useful, first for the estimation of their statistical information content, and second as an aid in the design of coding systems. Some results of these investigations are given.

Choice of Statistic to be Measured

There are about 200,000 picture elements in a standard television frame, and if 32 brightness levels are allowed, then $32^{200,000}$ different pictures are possible. This number at once points out the unnecessarily high capacity of the existing television system and the effort that would be required to determine the complete statistical description of television signals. Such a description would entail the measurement of the probability of occurrence of each possible picture, individually and in sequence. Clearly a less complete statistical description, nevertheless useful, must be found. The choice is wide, and a selection is dictated by the use to which the data will be put.

Inspection reveals that the principal redundancy in television signals, and consequently the source of channel economies, lies in the relation among the amplitudes (or brightness) of neighboring picture elements within a frame, and among corresponding elements in successive frames. Correlation of this type has already been exploited to reduce the average power required for picture transmission.[2] This leads us to believe that the appropriate statistic to measure is the joint amplitude probability distribution of these related picture elements. An nth order distribution permits the calculation of an nth order upper bound on the information content. This in turn sets a limit on the maximum bandwidth compression possible using a coding system which utilizes just the statistical relation among the neighboring elements. If, in fact, a significant statistical relation existed among, say, five such elements, we would want to measure a fifth order distribution. If we measured a lower order distribution, we would then calculate a looser (higher) upper bound on the information content, and would be more pessimistic than necessary about the prospects for bandwidth reduction.

The fact is that we do not know how many

nearby elements are statistically related. Therefore we shall measure as high an order distribution as present techniques permit, and compare the results with lower order measurements. That is the principal reason why we have chosen to measure third order distributions. However, there is an additional reason which may be more compelling. This is that it appears to us that a practical coding system (at least the first practical coding system) ought to use the same code for all pictures. To do otherwise would require a great deal of additional equipment. Therefore the statistic worked on ought to be one which is reasonably constant for the range of pictures encountered in television practice. There is a plausible argument that the third order distribution is the highest such. The argument is this: For a picture of 64 allowable brightness levels, a third order distribution is a tabulation of the probabilities of occurrence of each of 64^3 combinations of brightnesses. This is about one quarter million combinations. Since there are only about 200,000 elements per frame, each of these combinations will occur on the average about once per frame. Due to the nature of pictures, most of these combinations will never occur, and some, for example, those indicating three equal brightnesses, will occur rather frequently. However, there will be some combinations which occur in typical pictures once or twice. Obviously, a reliable estimate of these probabilities is not obtained. The measurement of fourth or higher order distributions would result in a much larger proportion of such combinations in which the statistical evidence was very unreliable.

Finally the third order distribution will actually give us a great deal of information. It will, for example, settle once and for all the question of whether the generation of picture signals is a second or third order statistical process. For example, if the only statistical influence which exists in a picture is between the brightness in adjacent elementary areas, then the third order approximation to the entropy will be equal to the second order approximation and this is a point which can be resolved by this measurement.

Technique of Measurement

The amplitude probability distribution of a stationary time series (i.e., a sequence of numbers or pulses) is measured by means of counting the number of occurrences of each possible value of the variable for a long time and normalizing by dividing by the total number of pulses or numbers. For a periodic time series, it is necessary to count only for one period. The distribution of a continuous variable may be measured in two ways. The first way is to measure the proportion of time during which the signal is found within each small amplitude interval, for example by blackening a photographic emulsion in proportion to these times, [3,4] or by averaging a current which flows only when the signal is in the interval. This is the method used previously by the author in measuring second order distributions.[5] The advantage of the measurement by proportions, which may be called a continuous measurement, is that the equipment can be relatively simple. In addition, it is generally possible to make a simultaneous measurement of the probability for all of the separate amplitude intervals, and thus obtain the entire distribution very quickly. On the other hand, this method is also characterized by a limited dynamic range. The ratio of maximum to minimum probability which can be measured reliably is ordinarily limited to about 100:1 by stray light and/or the characteristics of photographic materials.

A second way to measure the distribution of a continuous variable, assuming it is band-limited, is to create a time series from it by sampling at a rate high enough (twice the highest frequency component) to include all the significant fluctuations, and then to measure the distribution of the time series by a counting technique. The advantage of this digital method is that, unlike the continuous technique, the dynamic range of the measurement is unlimited by unintentional factors, and is set solely by the number of events (picture elements) in a period. On the other hand, the circuits required to produce and process the sampled video signals are complex, and the counter, which for television signals must operate at a maximum rate of 8 Mc/sec., is an additional complication not present in the continuous case. Furthermore, unless there is provided a whole set of expensive counters, it is necessary to use one period to measure the count in each amplitude interval, or, in our case, for each of the 64^3 combinations of amplitude intervals.

Despite the added complexity of the digital measuring method, it was selected on account of its higher accuracy. There seemed to be no advantage in using multiple counters to speed up the recording of data, as only about 2 1/2 hours is required to record an entire distribution. Each count is recorded on magnetic tape for later use directly as the input to an IBM 704 computer for calculation of the entropy.

Description of the Apparatus

The probability machine consists of two basic parts. The first is a flying spot scanner which generates stable, high quality video signals from slide transparencies. The other part consists of the circuits necessary to measure and record the probability densities of the video signals produced by the scanner.

The key operations in the measurement of the probability distribution are sampling the signals at 8 Mc/sec, and then determining in which of the 64 amplitude intervals each sample belongs. Both of these operations are performed in a simple and elegant manner by a special cathode ray tube, called a switch tube, the design of which was evolved jointly with the tube manufacturer. Figure 1 is a drawing of the final model of the device. Because of the key role played by the switch tube, it is worthwhile discussing its characteristics in some detail.

The Switch Tube

Many investigators have used cathode ray tubes for probability measurements. The signal is ordinarily applied to one set of deflection plates and a measurement is made of the brightness of the trace at each point along its length. If the phosphor brightness is actually proportional to the average current density of the incident beam, then the brightness at each point is proportional to the length of time spent by the beam in the vicinity of each point, as desired. In gray wedge analysis, the brightness is measured photographically. It is possible to eliminate defects in this measurement due to nonlinearity of the CRT deflection and due to non-uniformity of the phosphor by measuring the brightness photoelectrically and translating the trace past a fixed aperture, rather than vice versa.

It appeared to be an improvement over this last method to use a special cathode ray tube which eliminated the steps of converting electricity into light and back into electricity by permitting the direct measurement of the beam current through a physical aperture corresponding to the aperture of the optical arrangement. This eliminates any error due to phosphor saturation, light scattering by the phosphor, and stray light. However, the dynamic range of the measurement is still limited by stray electrons due to reflection and secondary emission. It is this consideration which led us to the decision to count pulses rather than to measure currents or brightnesses. This is done by applying to the cathode of the switch tube an 8 Mc/sec sampling pulse train. Then whenever the beam is directed towards the aperture, a train of current pulses is produced in the collector electrode, and these may be counted. For determining the amplitude interval in which a signal belongs, the voltage applied to the deflection plates is made equal to the difference between the video signal and a fixed reference voltage. Since the aperture is centered at zero deflection voltage, an output will be obtained whenever the video signal is equal to the reference voltage, plus or minus the amount required to deflect the beam to the edge of the hole. By setting the hole size to correspond to one sixty-fourth of the peak-to-peak signal amplitude, each signal sample will in principle produce an output pulse for one and only one of the sixty-four possible values of the reference voltage. By having the reference voltage take on each of its possible values for one frame interval at a time, and by recording the number of pulses through the aperture in each frame, a complete first order distribution is recorded in 64 frames. A second order distribution may be measured by applying a second video signal and reference voltage to the other set of deflection plates of the switch tube. Under those conditions, an output will be obtained only when both video signals match their respective fixed voltages, the switch tube simultaneously performing the functions of amplitude selection and coincidence. The second fixed voltage then is made to change one step each time the first fixed voltage completes an entire cycle of 64 steps. 64^2 or 4096 frames are used for the second order measurement. Similarly, 64^3, or 262,144 frames are required for the third order distribution. For that measurement, a second switch tube is needed, and a coincidence must be formed between the outputs of the two tubes in advance of the counter.

The ideal operation of the switch tube as described above depends on having an electron beam of perfect focus, i.e., zero diameter. The effect of a finite diameter beam is to produce collector current pulses of intermediate amplitude when only a portion of the beam enters the aperture. To minimize the proportion of pulses which are so affected, it is desirable to have as high a ratio as possible of aperture width to beam diameter. Since the aperture width corresponds to 1/64 of the video amplitude, we use as high a video voltage as convenient to generate and use a physical aperture of corresponding size. However, once one has set a limit to the deflection voltage, it is still possible to select operating conditions for the switch tube to maximize the ratio. The figure of merit of the tube for this purpose is the deflection sensibility, i.e., the voltage required to deflect the beam one beam diameter. This is so because, having set the amplitude of deflection voltage, we have also set the voltage required to deflect the beam one aperture width. Therefore the desired ratio is maximized by minimizing the voltage required to deflect the beam one beam diameter.

We have measured deflection sensibility and have found that it increases with decreasing acceleration voltage in the gun of the tube. Apparently, the deflection sensitivity increases faster than the beam diameter, as the acceleration is lowered. The disadvantages of low voltage operation are that the available beam current is low and the beam is susceptible to hum field deflection, but these difficulties have been overcome. We operate with about 500 volts acceleration, 10 microamps beam current, and have a close-fitting magnetic shield over the entire tube. With these operating conditions and a deflection voltage of about 200 volts peak-to-peak, plate-to-plate, the beam diameter is about 1/5 the aperture width of .25 inches.

One final consideration in the application of the switch tube concerns the shape of the current pulse in the collector. If the video deflection voltage is not changing during the interval in which the switch tube is pulsed on, then the current pulse has the same shape and duration as the 8 Mc/sec sampling pulse, which is about 1/16 μsec wide. If, however, the beam moves past the aperture in less than 1/16 μsec, then the current pulse shape and duration will be governed by the sweep speed, and, furthermore, pulses will be received by the collector for several successive values of reference voltage, since, in the same 1/8 μsec period, the video voltage will be in several adjacent amplitude levels. Whether these pulses are counted depends on their width and amplitude, and is thus uncertain. To overcome these difficulties, the video signals are processed by "boxcar" circuits which introduce steps into the signals to hold their

levels approximately constant for the active period of the switch tubes.

The operation of the probability machine can now best be understood by reference to the block diagram, Figure 2. It is convenient to consider first the generation of signals by the flying spot scanner and then the measuring and recording of the probability distribution by means of the switch tubes and associated circuits.

Flying Spot Scanner

The function of the scanner is to produce three separate video signals from three transparencies. Usually the three transparencies will be identical and the three signals will be in effect derived from three spatially related spots scanning one picture. To achieve this result, the blank raster of a high intensity flying spot cathode ray tube is imaged by means of a projection lens and two beamsplitters onto three glass transparencies. The light transmitted by each transparency is collected by condenser lenses and spread evenly over a portion of the photocathode of a photomultiplier tube. This is shown schematically in the block diagram. Figure 3 shows the entire optical system. The scanner tube is at the left. The projection lens, which is mounted on the end of the beamsplitter housing, is partially visible thru the open access door. The first beamsplitter reflects one-third of the light to the right, the remainder continuing on to the second beamsplitter, which reflects 1/2 up and permits the rest to pass thru. The straight-through and right-hand beams illuminate transparencies in movable holders, while the holder in the vertical beam is fixed. Each transparency is followed by condenser lenses and a phototube.

In a properly designed flying-spot scanner, the main source of noise is shot effect in the photocurrent. Since the noise voltage is proportional to the square root of the current, the signal-to-noise ratio is proportional to the square root of the current. Noise is thus reduced solely by increasing the photocurrent, i.e., by increasing the light incident on the phototube and by using the most sensitive photocathode available. We have attempted to meet these goals by the following methods:

1. The combination of P16 phosphor and S4 photocathode is the most efficient known.
2. The scanner tube operates at the highest voltage and current ratings in current engineering practice.
3. The objective lens is the widest aperture lens available of this focal length (selected for other reasons) having adequate definition.
4. The semi-reflecting mirrors are interference beamsplitters of very high efficiency.
5. The condenser lenses are made of Pyrex glass having good transmission in the spectral region of interest and are coated.
6. Since for a given effective f number, the brightness of an image is independent of its size, the total light energy incident on an image is proportional to its area. Hence the largest practical transparencies, i.e., 3 1/4 x 4", with a useful area of 2 1/4 x 3", have been used, resulting in a signal more than 4 times larger than if Leica size transparencies had been employed.

Further precautions are necessary to ensure that the three images are of equal size and high quality. Each beamsplitter is mounted on a very thin, flat support, called a pellicle, which consists of an organic membrane stretched tightly over a carefully lapped frame, in the manner of one-shot color cameras. The frames themselves are rigidly mounted in an accurately machined box.

Glass transparencies are used for dimensional stability. All in a set are exposed one after the other in an electronically timed high quality enlarger, and developed at the same time in a hand-agitated rack so designed that each of the plates undergoes the same development.

In the optical system the two adjustable slide holders are movable in such a manner that no difficulty is encountered in registering the three pictures. A sensitive means of determining when registry is achieved is to superimpose the signals, in pairs, on a picture monitor. Micrometer adjustments, which can be seen in Figure 3, allow 2 of the images to be displaced horizontally or vertically with respect to the third by measured amounts to derive the three signals needed for the probability measurement.

In addition to the purely optical method of deriving the two additional signals needed for the measurement, delay lines of 1, 2, and 3 Nyquist intervals (1/8 μsec) are used, so that there are a total of six different signals available for comparison.

Amplifiers

The photomultiplier outputs are amplified and equalized for phosphor persistence and limited to 4 Mc/sec bandwidth in the preamplifiers ("P" in Figure 2). The signals are then passed to the distribution and delay circuit where the three signals for analysis are selected from the six signals available. The three deflection amplifiers raise the signal level to the value required for the switch tubes. The monitor amplifier permits the observation of these three signals, in pairs, on the face of the picture monitor, as an aid in setting up the equipment. There is also a circuit available for taking cell-to-cell difference signals. It uses a shorted delay line of one Nyquist interval round trip length, the line being terminated at its sending end in its characteristics impedance, so as to add to the input signal the inverted, delayed, returning signal.

Another circuit not shown rectifies the deflection amplifier outputs and feeds control signals back to the photomultiplier tube power supplies for the purpose of holding the signal level constant at the switch tubes, even though the scanner tube brightness, the phototube sensitivity, or the amplifier gain should change. This arrangement holds the amplitude constant to within several per cent with a ten-fold change in gain anywhere in the system. At the same time it eliminates the need for closely regulated photomulti-

plier power supplies.

Coincidence Amplifier and Counter

Each switch tube output is amplified and applied to a diode coincidence circuit and then amplified again sufficiently to operate the first stage of a 17-stage binary counter. The switch tube output is adjusted by controlling the beam current so that when the beam is centered on the edge of its aperture, giving an output signal one half that obtained when the beam is completely in the hole, the signal at the counter input is just enough to operate it. In this way, a count is recorded whenever the center of the beam is within the aperture.

Each counter stage except the first is capable of operating at least twice as fast as the maximum rate at which it is called upon to operate in use. The maximum rate for the first stage is 8 Mc/sec, and it is capable of 12 Mc/sec operation. Cathode-follower coupled triodes are used in the first two stages, followed by 6 stages of amplifier-isolated triodes, and finally 9 medium power dual triodes. 17 stages are necessary only for pictures which are almost entirely a single brightness level, in which case a count is recorded for each picture element in the one-brightness area.

Staircase Generators

These three circuits generate the 64-step reference voltages which are applied to the switch tubes along with the video signals. Each one includes a 6-stage binary counter. The last stage of the first staircase generator feeds the first stage of the second generator and the last stage of the second generator feeds the first stage of the third generator. As shown in Figure 4, the staircase voltages are produced by having each "high" stage in turn cause a pentode to draw a carefully fixed current through a precision voltage divider. By use of very high plate resistance pentodes, optimum operating point, a large amount of degeneration, drastic derating of components, push-pull operation, .05% resistors, and separately regulated plate and screen voltage supplies, a linearity and long-term stability of about 1 part in 1000 is achieved. "Utra linear" cathode followers of a special design are used at the staircase generator outputs so that large capacitors may be driven, providing a low impedance bias for the switch tube clamping circuits. These cathode followers have a performance matching that of the staircase generators and provide low impedance outputs for both positive- and negative-going signals of large amplitudes.

Readout Circuit and Recording Amplifier

The seventeen stage binary counter which records the number of occurrences, in a frame interval, of the particular combination of video brightnesses being measured, receives no counts during the vertical blanking period, since the switch tube is cut off. During this interval, the readout circuit interrogates the counter, stores the seventeen digits on 17 storage capacitors, and then resets the counter. During the ensuing frame, the stored bits are read out, six at a time, and passed to the recording amplifier for permanent storage on the six information tracks of the magnetic tape. The seventh track on the tape is used for timing, and is supplied, by the readout circuit, with a continuous 90 cycle signal coincident with the information pulses, if they are present.

Recording is by the non-return-to-zero method, in which the tape is saturated in the track, the direction of saturation being changed when it is desired to record a one, no change being made for a zero. This style of recording is reliable and easy to read, and produces easily visible marks when developed in a suspension of magnetic particles, as shown in Figure 6. (Commercially available as Ferroprint Magnetic Inking Solution)

The format on the tape is exactly that necessary for the tape to be used as the input to the IBM 704 computer. Our tape recorder moves continuously at .45 inches per second, thus recording 200 bits per inch in each track.

Clock Pulse Generator

This circuit generates a train of 8 Mc/sec pulses, starting at each horizontal synch pulse, and continuing throughout each TV line. It consists of a ringing circuit with feedback to keep the output constant. The pulses are used for sampling at the switch tube and also in the "boxcar" circuits previously described.

Cycling Circuit

Because of the fact that data must be taken so fast and recorded as it is taken, it is necessary that all functions in the apparatus be completely automatic. The cycling circuit controls all the other circuits for the 2 1/2 hour data recording period, and turns off the appropriate signals at the end.

Before starting, the switch tubes are clamped off at their control grids, so that no pulses are transmitted to the counters. Clock pulses, insufficient to turn the tubes on, are continuously applied to the switch tube cathodes, and the video signals are continuously applied to their deflection plates. When data recording is to start, the tape is set in motion, the staircase generators are manually reset, and the "start" button on the cycling circuit is pressed. The next following 30 cycle pulse, obtained by division from the 60 cycle vertical blanking signal, opens a gate, passing blanking pulses to the switch tube and thirty cycle pulses to the first staircase generator. When, 2 1/2 hours later, the third generator returns to its "low" position, the gate is closed and recording stops. During this period, the blanking signal turns the switch tube on for the peak of each clock pulse which occurs during every active horizontal line. In addition, various spaces and extra marks are provided on the

tape, as required by the computer.

The thirty cycle pulse is also sent to the readout circuit, where it initiates the train of events described previously.

Figure 5 is a general view of the apparatus: The tape recorder is in the right hand rack, on slides, mounted beneath the recording amplifier. The next rack contains the three deflection amplifiers, with meters to indicate output level. Immediately above is a monitor scope connected in parallel with the first switch tube, for the visual observation of second order distributions. The three lowest units in the next rack are the staircase generators, above which are the cycling and readout circuits. The preamplifiers and picture monitor are on the shelf above the optical system, while the space below the optical system is occupied by power supplies.

The Computation

The data taken are actually the number of occurrences, in a frame, of the particular combination of three brightnesses corresponding to the three staircase generator settings during that frame interval. If these data are divided by the total of all occurrences of all combinations in a run, they are true probabilities. That is, if n(i,j,k) is the datum when the staircase generators are on their ith, jth and kth step, respectively, and p(i,j,k) is the probability of this combination, then

$$p(i,j,k) = \frac{n(i,j,k)}{\sum_{i=1}^{64}\sum_{j=1}^{64}\sum_{k=1}^{64} n(i,j,k)} = \frac{n(i,j,k)}{N} \qquad (1)$$

where we define N as the total of all the counts of a run. N is equal to the number of Nyquist intervals in a frame, or about 190,000. The third order joint entropy, i.e., the information generated by three picture elements, is

$$H(x,y,z) = -\sum_{i=1}^{64}\sum_{j=1}^{64}\sum_{k=1}^{64} p(i,j,k) \log p(i,j,k) \qquad (2)$$

where H is in bits if the logarithms are binary. Omitting indices for clarity,

$$H(x,y,z) = -\Sigma\Sigma\Sigma \frac{n}{N} \log \frac{n}{N} = \qquad (3)$$

$$-\frac{1}{N} \Sigma\Sigma\Sigma (n \log n - n \log N)$$

$$H(x,y,z) = \frac{1}{N}\left[N \log N - \Sigma\Sigma\Sigma\, n \log n\right] \qquad (4)$$

which is a form particularly suited for computation. The information on the tape is divided into records 2048 numbers in length, corresponding to 32 cycles of the first staircase generator. Thus a second order distribution fills just two records and a third order distribution fills 128 records. The numbers are put into storage in the computer, one record at a time. As the data enter storage, the sum of n is accumulated over each 64 values and over the record. The sum of n log n is accumulated over the record, the value of n log n being found from a stored function table containing the first 1024 values. After the record is in storage, numbers larger than 1024 are detected by comparing the sum of n over the record in storage, effectively using only the first ten digits of each datum, with the running sum. Most records do not contain values higher than 2^{10}, in which case the two figures match. If they do not match, then each value of n is tested to find those greater than 1024. Large values are computed using a scale-changing formula and interpolation, and the correct value of n and of n log n then effectively replace the incorrect values in storage, the sums being recomputed.

The large-number procedure is also used to compute the 64 values of X log X for each pair of records, the X's being the partial sums of n over 64 values, i.e.,

$$X(j,k) = \sum_{i=1}^{64} n(i,j,k) \qquad (5)$$

The 64 X log X are summed and stored, the X's themselves being discarded. Finally, the large-number procedure is used to calculate the one value of Y log Y for the pair of records, Y being the sum of n over the two records, i.e.,

$$Y = \sum_{j=1}^{64} X(j,k) \qquad (6)$$

At the end of each pair of records, we store the sum of n log n, the sum of X log X, and the one value each of Y and of Y log Y. At the end of the file, we print out these results for each of the 64 pairs of records, and in addition, calculate and print out N, the sum of n over the file and the sums of n log n, X log X, and Y log Y over the file. The one value of N log N is hand-calculated for each file.

It is evident that the X's constitute a second order marginal distribution and the Y's constitute a first order marginal distribution. It is useful, therefore, to calculate the corresponding first and second order entropies:

$$H(z) = \frac{1}{N}\left[N \log N - \sum_{k=1}^{N} Y \log Y\right] \qquad (7)$$

$$H(y,z) = \frac{1}{N}\left[N \log N - \sum_{j=1}^{64}\sum_{k=1}^{64} X \log X\right] \qquad (8)$$

The most significant entropies are $H_{y,z}(x)$ which is the information generated by the single picture element x when the values of two other elements, y and z, are known, and $H_z(y)$ which is the information generated by one picture element when the value of one adjacent element, z, is known. By Shannon's formulae,

$$H_{xy}(z) = H(x,y,z) - H(x,y) \qquad (9)$$

$$H_y(z) = H(y,z) - H(y) \qquad (10)$$

and these are found by subtraction of the entropies previously calculated. Note that the indices may easily be shifted around provided the relative displacements of the original scanning apertures are preserved.

Measurements

Second Order Measurements

The second order distributions of a number of subjects of varying complexity have been measured. In each case, our procedure has been to record three distributions for each subject, with the two video signals in register and then displaced first one and then two Nyquist intervals (1/8 μsec). Results for two subjects are tabulated. A was our most complicated and B our least complicated picture.

TABLE I

Subject	Displacement	H(x,y)	H(x)	$H_x(y)$
A	0	7.35	5.65	1.70
A	1	9.06	5.70	3.36
A	2	9.77	5.73	4.04
B	0	5.27	4.32	0.95
B	1	6.15	4.30	1.85
B	2	6.64	4.35	2.29

The maximum possible value for H(x,y) is 12 bits per symbol, and this would occur only in a completely random picture of flat amplitude distribution. The maximum value of H(x) is 6 bits per symbol and this would occur only in pictures of flat amplitude distribution. $H_x(y)$ is mostly a measure of the intersymbol correlation, the effect of the first order distribution having been largely discounted by the conditional (i.e., previous value known) nature of this parameter. The minimum value of the conditional entropy is zero, a situation which occurs only when each picture element is exactly like the previous one. In terms of the scatter pattern which develops when the two signals are applied to the two sets of deflection plates of the switch tube and monitor tube in parallel, we would obtain a thin diagonal line, and when the pattern was quantized for measurement by the staircase generators, only the 64 diagonal elements in the array would give non-zero counts. Now in practice, due to non-linearity, noise, voltage drifts, and the like, we actually record counts in somewhat more than 64 elements. If, for example, all the counts that should be counted in one square are instead spread over two squares, the measured conditional entropy is one bit per symbol. Now as this is quite a small displacement, amounting to a noise of something like 1/64 the amplitude of the signal, it can easily be seen that when we calculate for any picture a conditional entropy around one bit, what we obtain is not the real information content of the picture, but a measure of the difficulty of producing and processing video signals to that accuracy, at the present state of the television art.

In our previous work at Harvard, the minimum measured entropy for a 32 level picture was about two bits. As this was largely attributable to noise, we devised an approximate method of correcting the measured results to eliminate the effect of noise. In the present work, where we measure minimum entropies of about one bit with a 64 level picture, we have decided against making any noise correction. We now prefer to look at the coding problem in terms of dealing with actual pictures, of high but realizable quality. It should be emphasized, therefore, that the figures in the table above and elsewhere in this paper, are actual experimental data, and are not corrected in any way for the deviation of the signals from ideal.

Another point worth noting about the data in the table is that H(x) is the entropy of the first order marginal distribution. As such, it should not vary as the displacement between the signals is changed, and in fact it is usually constant to within several percent, although naturally it varies considerably from subject to subject, depending on whether the pictures have a broadly distributed range of brightnesses or are mostly dark or light.

The entropy most important for coding is the information generated by one picture element when the brightness of the adjacent element is known. These values are underlined in the table and vary from 1.85 to 3.36 bits per symbol. The average of all the subjects we have measured is 2.62 bits per symbol. From these results, it would seem that the amount of bandwidth compression obtainable using only adjacent cell correlation is quite low, and that other relations must be investigated if a really efficient system is to be found. To this end we have also measured the second order distributions of difference signals which are obtained by subtracting from each picture element the brightness of the previous element. Although this linear operation does not affect the entropy of the picture if all the picture statistics are used, it may conceivably affect the low orders of approximation to the total entropy which we are measuring.

A complication in this measurement is that our equipment handles just 64 brightness levels, but a difference signal has twice as many levels as the original, since differences may be both positive and negative. Therefore, in each case where we desired to measure a difference signal, we first measured the original distribution with the amplitude reduced to cover only 32 of the 64 levels available. The resultant entropies are of course lower than those of the full amplitude signal, because the distribution is less uniform. We have found that H(x) goes down something less than one bit, and $H_x(y)$ a little less than that. Where the conditional entropy is already very low, it decreases hardly at all. When differences are now taken on this half amplitude signal, a full amplitude signal again results, the first order distribution of which, as expected, is sharply peaked at mid-amplitude, representing no change from cell to cell. The first order entropy usually is about one bit less than that of the half amplitude signal; the conditional entropy of the difference

picture may be only slightly lower than the conditional entropy of the original, or it may be very much lower, and we have not measured enough pictures of this type to be able to give a final result. A promising feature is that in many pictures, differencing appears to reduce the first order entropy to almost as low a value as the conditional entropy of standard pictures. If further measurements bear this out, it means that coding on a symbol-to-symbol basis can be made almost as efficient as second order coding, by means of differencing, with a significant saving in equipment.

Third Order Measurements

We have also measured a number of third order distributions. It will be recalled that our computation provides for the derivation of the second and first order marginal distribution and the calculation of their respective entropies, as well as the calculation of the third order entropy. It is a valuable checking procedure, therefore, to compare the second order results from third order measurements with the second order results from second order measurements. A typical result is tabulated below, where the subject is B, our least complicated picture, and the scanning apertures are horizontally disposed and separated by one Nyquist interval spaces.

TABLE TWO

Measurement	$H(x,y,z)$	$H_{xy}(z)$	$H(x,y)$	$H(x)$	$H_x(y)$
2nd order	--	--	6.15	4.30	1.85
3rd order	7.80	1.49	6.31	4.39	1.91

The second order entropies in the two measurements are within about 3% of each other. Considering that the runs in question were made at different times and that the equipment was moved to a new laboratory in the time between, this is reasonable precision. Another point worth noting is that $H_{xy}(z)$, the information generated by one picture element when the brightness of the two preceding elements is known, is significantly lower than when the brightness of but one preceding element is known. This indicates that picture generation is in fact a higher-than-second order process. However, the amount by which the information content is reduced by the extra condition, though significant, is small. Unless future measurements show a decidedly greater reduction, it is a reasonable conclusion that the point of diminishing returns has been reached in this direction and that a coding system, based on third order statistics of horizontally disposed picture elements, would not produce sufficiently greater compression than a second order system to warrant its greatly increased complexity. This result is in accord with (but could not have been predicted from) Harrison's[2] experiments in which he reduced the average power of transmission by linear prediction and found that slope prediction, where the two preceding elements were taken into account, was not substantially better than the simple prediction that each element should have the same brightness as the one preceding.

Typical Distributions

In addition to having our data computed, it is also possible to have it printed out. At first thought it might seem that the volume of data would be so large that this would be a very awkward procedure. However, most of the combinations of brightness studied never exist in the picture and so much of our data is zero. The print out procedure skips blocks of zeros, thus compressing both the time required for print out and the bulk of the paper.

In Figure 7, we have plotted three (of the 64) "cuts" through a second order distribution. They are each sharply peaked at j=k, where the two brightnesses are equal. The curves are generally bell-shaped and it is characteristic that the width of all the cuts in a pattern is about the same. In Figure 8 we have hand-calculated the two first order marginal distributions from a printed out second order distribution, to check whether they matched as they should. In fact, they match quite well except for what seems to be a slight amplitude non-linearity in one of the signals, which would not have much effect on the calculated entropy. Figure 9 shows a few cuts through a third order distribution. These have much the expected shape. The fact that the peak is not exactly at i=j=k is due to a slight difference in amplitude among the three signals and again would have little effect on the entropy. A feature of these curves is the rapidity with which they decrease from maximum value for other combinations of brightnesses. It is also no accident that the peak of each curve for successive values of j nearly coincides with the height of the adjacent curve at the next value of k, since this is merely an indication of the symmetry of the distribution in j and k.

Figure 10 consists of photographs of the second order distributions whose entropies are given in Table 1. They spread out with successively larger displacements, as expected. The similarity of the two channels and the noise level may also be judged by observing these patterns.

Acknowledgement

Thanks are due to D. H. Kelly, who interested the company in this work, and to John R. Clark, Jr., Technicolor Vice President, for his support of the project. G. T. Inouye and C. O. Carlson assisted in design problems and the latter has been in charge of operating the machine. M. Reinhard did all the mechanical design. E. H. Borchardt, R. S. Hiatt, W. Cahn and W. Gismot constructed the electronic circuits. Frances House did computations, drew graphs, and typed the paper. A. E. Mann conceived the method of machine computation and did most of the detailed programming. We have received excellent cooperation from IBM personnel at all levels, and in particular from B. O. Evans, who helped with our tape problems, and Helene Steinman and Mrs. M. Levine, programmers, who had the additional job of guiding our "foreign" tapes through the intricacies of the Electronic Data Processing Machines.

References

1. P. Toulon, L'Onde Elect. Vol. 28, p. 412, 1948

2. C. W. Harrison, Experiments with Linear Prediction in Television, Bell System Tech. J., Vol. 31, No. 4, p. 764, July, 1952

3. E. R. Kretzmer, Statistics of Television Signals, Bell System Tech. J., Vol. 31, pp. 751-763, July, 1952

4. W. Bernstein, R. L. Chase, and A. W. Schardt, Rev. Sci. Instr., Vol. 24, No. 6, p. 437, June, 1953

5. W. F. Schreiber, Ph. D. Thesis, Harvard, 1953; IRE Convention Record, Part 4, p. 35, 1953

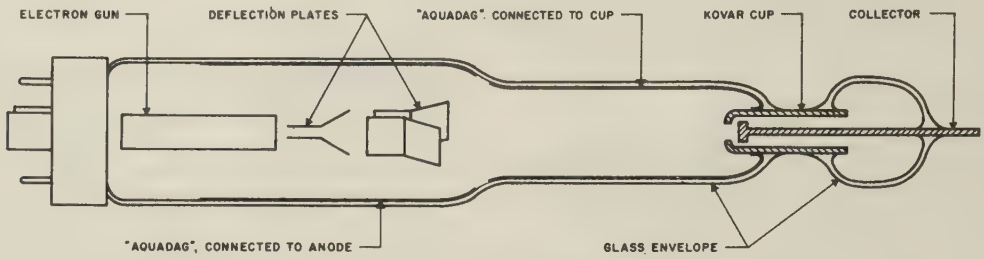


Fig. 1 - Switch tube.

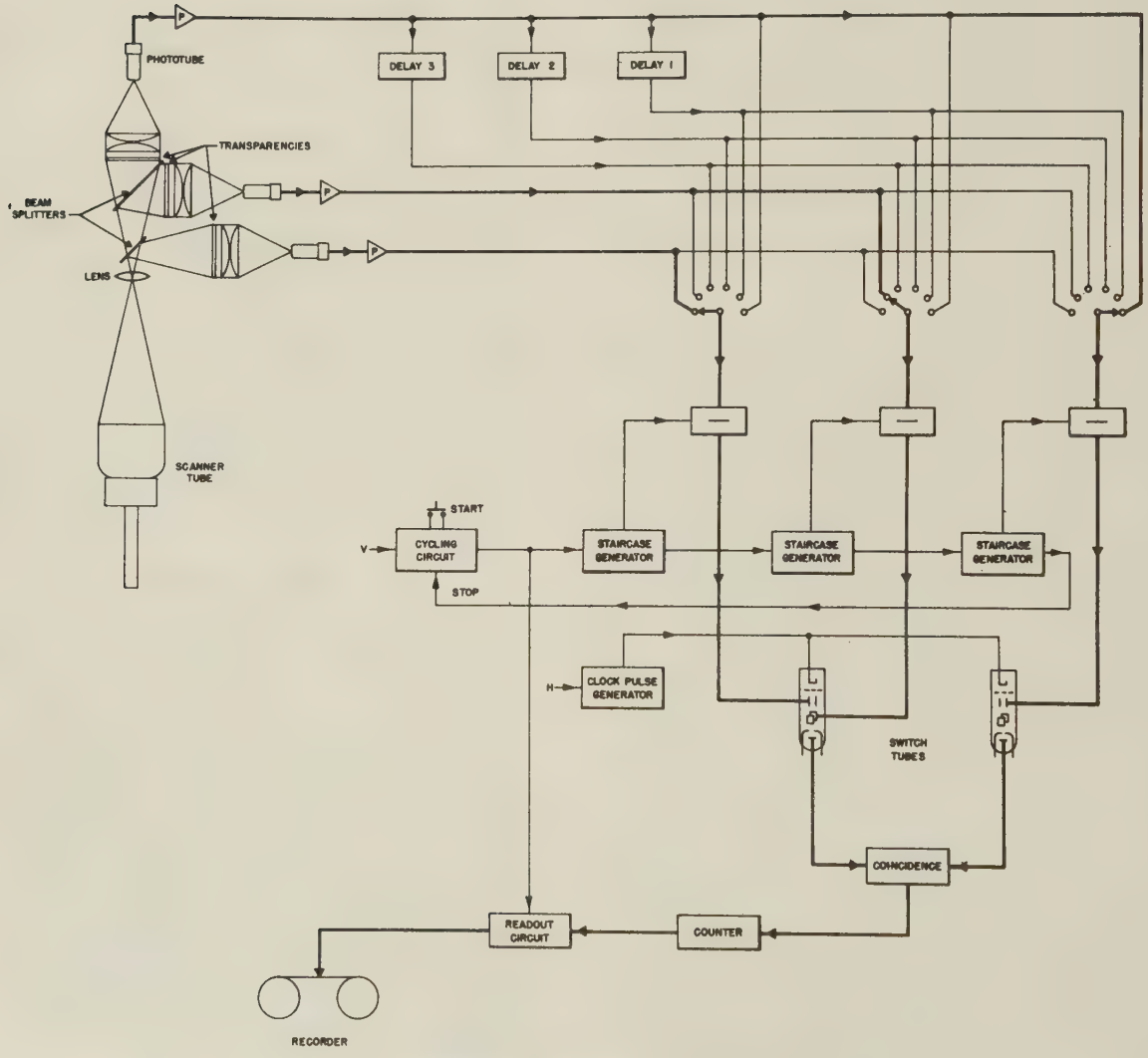


Fig. 2 - Probability machine -- block diagram.

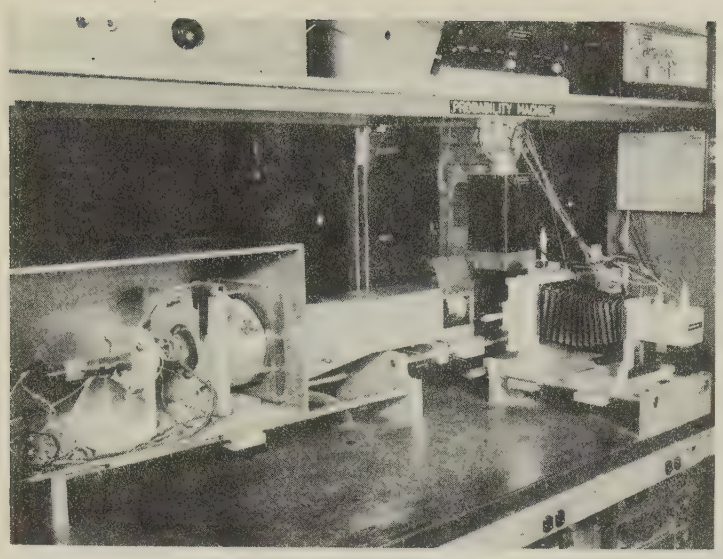


Fig. 3 - Optical system.

Fig. 5 - General view.

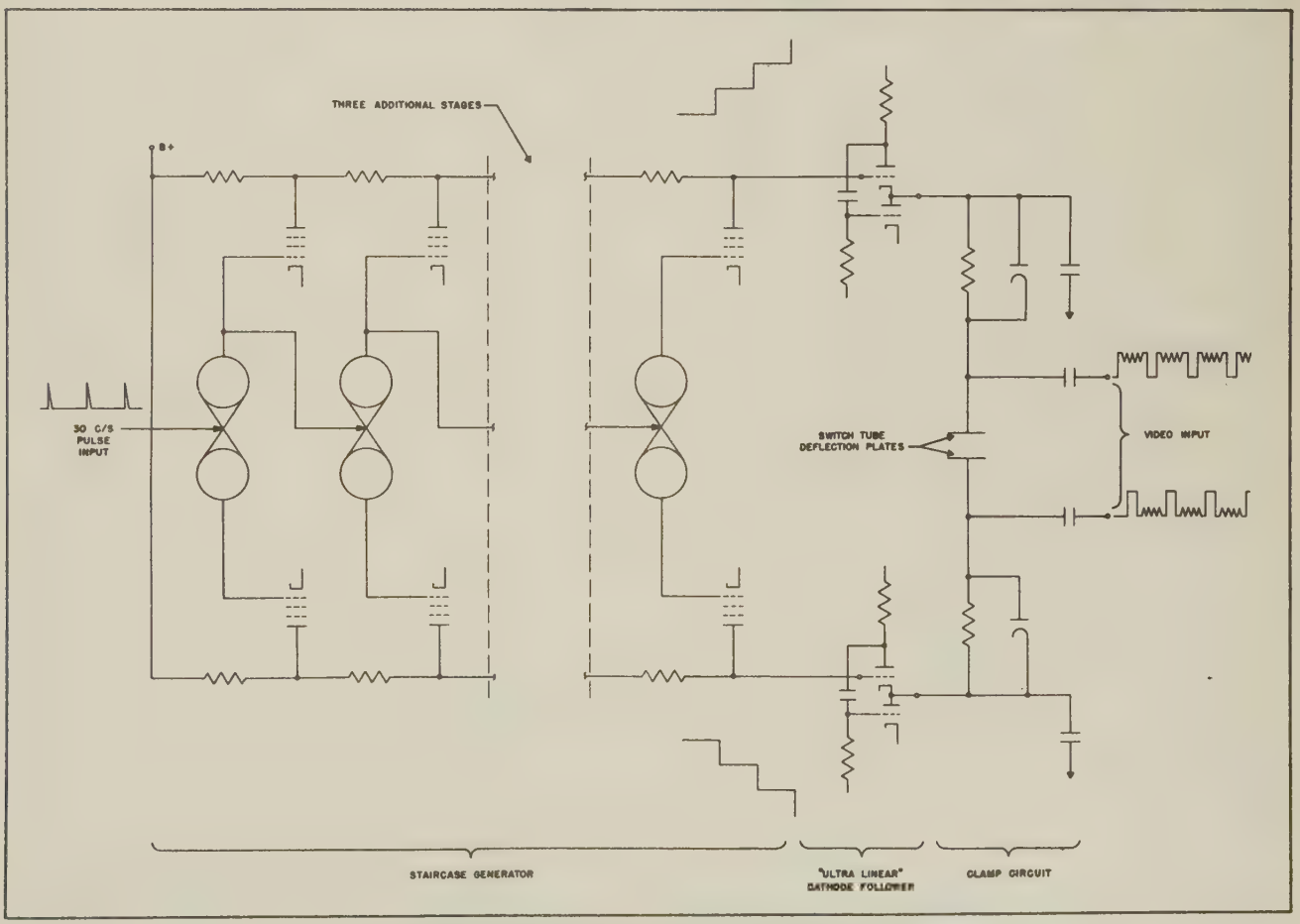


Fig. 4 - Staircase generator and video coupling arrangement.

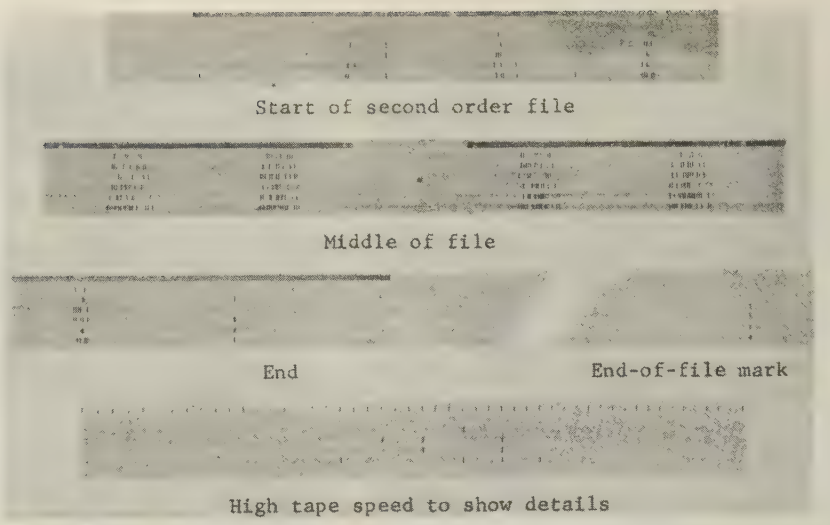


Fig. 6 - Developed tape record.

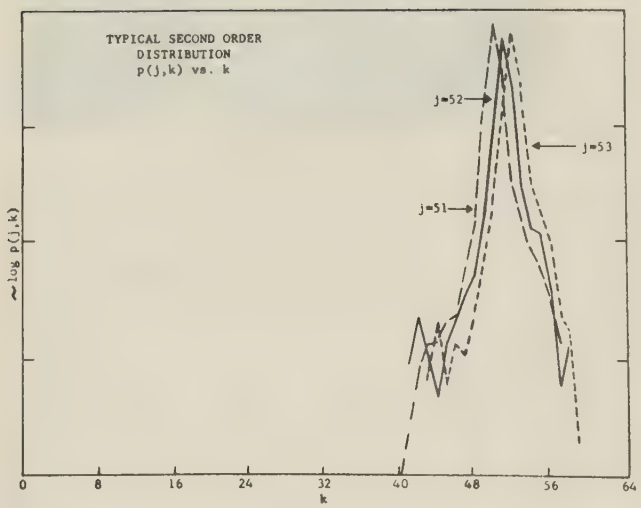


Fig. 7 - Typical second order distribution.

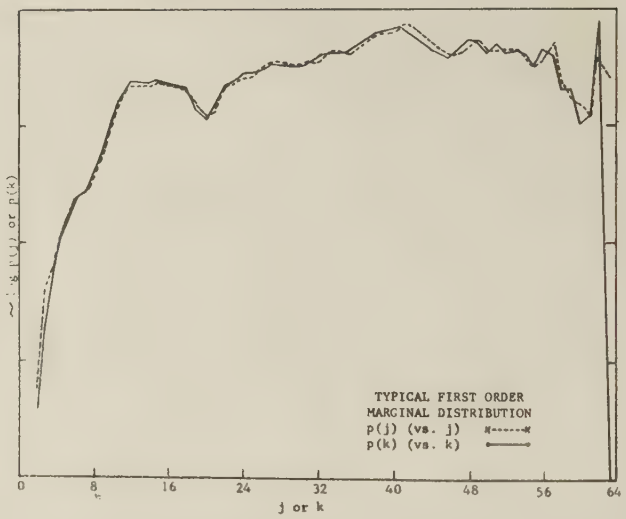


Fig. 8 - Typical first order marginal distribution.

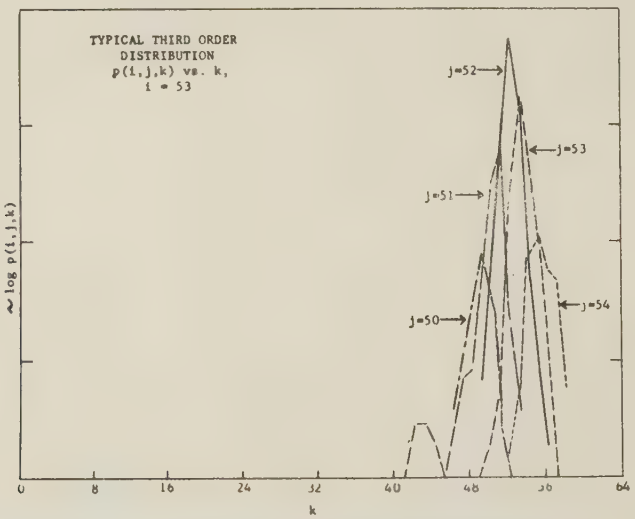
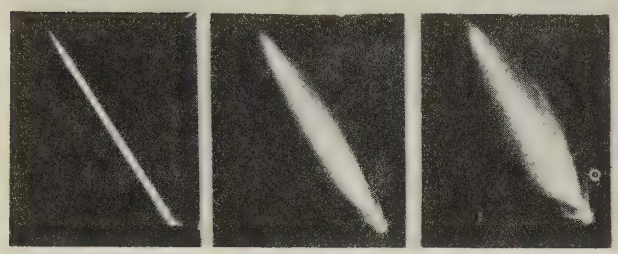


Fig. 9 - Typical third order distribution.

Subject A

Subject B



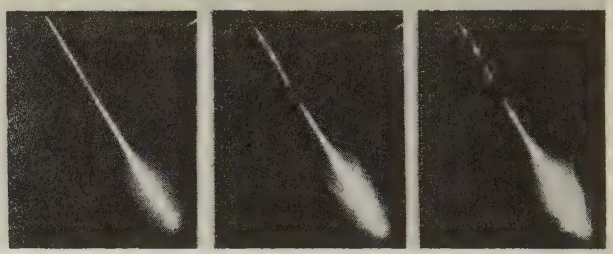


Fig. 10 - Second order distributions on the monitor scope. The simplest and most complicated subjects, together with second order distributions for displacements of 0, 1 and 2 Nyquist intervals.

GAP ANALYSIS AND SYNTAX*

Victor H. Yngve
Department of Modern Languages and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Summary

A statistical procedure has been tried as a method of investigating the structure of language with the aid of data processing machines. The frequency of gaps of various lengths between occurrences of two specified words is counted. The results are compared with what would be expected if the occurrences of the two words were statistically independent. Deviations from the expected number give clues to the constraints that operate between words in a language.

Introduction

Language is a very complex communication code. One of the tasks of the linguist is to discover the structure of languages, or the rules of the codes, and to state them in a simple and concise way. To do this, he collects for data actual samples of a language, looks for regularities in the data by applying various procedures of analysis and an appropriate amount of intuition, forms hypotheses, and tests them on more data. When he is finished, he has what he calls a description, or a grammar of a language.

Many of the difficulties that the linguist faces in his task of discovering and describing structure stem from the very complexity of the code and from the large amounts of data that must be examined. It has been suggested that modern data handling techniques using punched card machines or electronic digital computers might be able to overcome difficulties that arise from the sheer bulk of data. The purpose of this paper is to discuss a way in which this might be done.

Our procedure is a statistical one and makes use of the fact that order and disorder are in a sense complementary. Statistical independence implies lack of structure, and any deviations from randomness can be taken as an indication of structure. The procedure, therefore, has two parts: one part deals with the setting up of an appropriate statistical model of language; the other part deals with the deviations from randomness exhibited by language and their interpretation in terms of structure.

We assume that language can be represented as a sequence of symbols. These might be letters, phonemic characters, syllables, morphemes, or other elements. It is convenient if there is an operational procedure for segmenting the text into separate symbols. Such a procedure exists for words in conventional spelling; they are separated in a text by spaces or punctuation. For purposes of the example in this paper, we have adopted English words as the symbols.

The obvious first step in the statistical analysis of a text is to investigate the relative frequency of the different symbols. Counting the frequencies of words has been a favorite occupation for over 60 years and a considerable amount of data exists[1]. In 1928, E.U. Condon[2] observed that when words are ranked in order of decreasing frequency, the product of the frequency and the rank is approximately constant. Several explanations have been offered for this or other formulations of the word-frequency distribution law. Of these, two are especially interesting.

B. Mandelbrot[3] assumes that words are separated by spaces and are spelled with letters to which a cost function is attached. A message composed of words with the observed frequency distribution transmits the maximum amount of information in the sense of Shannon, compatible with a given average cost per word.

H.A. Simon[4] assumes a simple stochastic model. The probability that the next word to appear will be one of the words that has already appeared n times is set proportional to the total number of occurrences of words that have each appeared n times. There is a constant probability that the next word will be a word that has not already occurred. The observed frequency distribution agrees with the one that will keep the fraction of the words that occur n times approximately constant.

The observed distributions are nearly the same for all languages. If the assumptions on which these explanations are based are valid for one language, they are valid for all languages. The Mandelbrot explanation involves an economy argument; the Simon explanation follows if users of a language try to maintain the word frequencies that they observe.

After an investigation of the frequencies of individual symbols, the next step of a statistical analysis of a text is an examination of intersymbol constraints. These are of more direct concern to the linguist because they are different for different languages.

Intersymbol constraints have been investigated for various purposes. For cryptanalysis

* This work was supported in part by the Army (Signal Corps), the Air Force (Office of Scientific Research, Air Research and Development Command), and the Navy (Office of Naval Research); and in part by the National Science Foundation.

purposes, frequency tables of two-letter and three-letter sequences have been tabulated. For the purposes of estimating the entropy of printed English, Shannon[5] has used various methods of measuring the conditional probabilities that various letters will follow certain sequences of letters. The conditional probability concept has been used[6] as the basis for a model of a human being regarded as a talking animal. A grammar is conceived as an enormous array or matrix of the conditional probabilities that each morpheme in the language will be produced after a given sequence of morphemes. A scheme of this sort focuses attention on each position in a text and on the effect there of the immediately preceding one, two, three symbols, etc.

The method of investigating intersymbol constraints reported here is also concerned with the conditional probability of finding a given word at a certain position in the text. But instead of specifying the immediately preceding one, two, or more, text positions and investigating the effect on the probability of the words found there, we specify certain words or word combinations and investigate their effect as they are moved around in the vicinity of the given word. An advantage of this is that it allows more direct investigation of the effect of the occurrence of a word on the probabilities some distance away. It also allows easy and rather full investigation of the effects of the most frequent words first. Being most frequent, these words have an especially great influence on the grammar.

The Procedure of Gap Analysis

The statistical model that we use is a model for a text divided into symbols (words). We assume that the frequency f of each word W and the total number of words N, or the length of the text, are given as a result of direct measurement. We assume that the probability of occurrence of each word is equal to its relative frequency, $p(W)=f(W)/N$, and is therefore independent of its position in the text and of what words are nearby.

We look for deviations from the assumption that the probability of a word occurring is independent of the words in the neighborhood. To do this, we choose two different words and investigate their effect on each other's probability of occurrence. Or we can investigate the effect that a given word has on other occurrences of the same word. We define a gap of type A-B as the number of words intervening between an occurrence of A and a later occurrence of B. We can have gaps of type A-A between two occurrences of the same word. For each type of gap, we count the number of gaps of length 0, 1, 2, ... This can be done easily by machine by collecting one sample of text for each occurrence of A. All samples should be the same number of words in length and should have the occurrence of A at the same position. Then, for each of the other word positions, the number of occurrences of B in all the samples is counted. The results can be plotted as a histogram of the number of gaps against the gap length. Gaps of type A-B can be plotted on the right of the center of the histogram; gaps of type B-A on the left.

Several features of the histogram presentation of gap data should be noted.

1. If the probability of occurrence of B is independent of its position with respect to A, we expect the distribution of gaps to be flat except for statistical fluctuations.

2. The expected number of gaps of length n is then independent of n and can be calculated from the given frequencies:

$$f(G_n) = p(B)\ f\ (A) = \frac{f(A)\ f(B)}{N}$$

3. We ignore the effect that the ends of the text have in reducing the possible number of gaps. Such effects will be small if the gap lengths investigated are appreciably smaller than the length of the entire text.

4. A histogram with gaps of type A-B plotted on the left and gaps of type B-A plotted on the right is a mirror image of a histogram with B-A on the left and A-B on the right.

5. A histogram of gaps of type A-A is symmetrical about the center position.

6. If a gap had been defined as the number of words occurring between an occurrence of A and the first occurrence of B, the assumption of statistical independence would, of course, lead to an exponential distribution instead of a flat one, a fact that seems not to have been understood by various counters of gaps. We would have for the frequency of gaps of length n,

$$f(G_n) = f(A)[1-p(B)]^n p(B) = \frac{f(A)\ f\ (B)}{N}\ e^{-kn}$$

where $$k = -\ln[1-p(B)] > 0$$

For our purposes, gaps of the exponential type are not as convenient because they are harder to count by machine, require more calculating to obtain the expected number and the expected deviations, and because histograms with gaps of type A-B on the left and B-A on the right are not mirror images of those with B-A on the left and A-B on the right on account of the different exponentials.

Trial Application to English Structure

In order to be in a better position to assess the results in a first trial of the above procedure, we selected a small sample of a familiar language - English. An article of about ten thousand words from a popular magazine was chosen. Since this was a rather short article, only six of the most frequent words were investigated. The total number of words was counted as well as the frequency of each of these six words. These

numbers are tabulated below:

word	frequency
the	599
to	252
of	241
a	221
and	207
in	162
	1682
(number of words in article)	9490

It can be seen that these six words alone account for over 17.5 per cent of the occurrences of words in the text. Punctuation was ignored.

Using these six words, all fifteen of the type A-B gaps were counted, and the six of type A-A. The results of the gap counting are presented in Figs. 1 and 2. Along the abscissa of each histogram are plotted the various word positions. For example, in Fig. 1-g, the word "the" in all of the "the" samples of text is placed at the center position. The length of the bars of the histogram represents the number of times the word "a" appeared at the various text positions to the right or left of the "the". The numbers along the abscissa give the length of the gap or the number of words intervening between the occurrence of the word "the" and the word "a". The six histograms of the gaps between two occurrences of the same word are shown in Figs. 1-a to 1-f. In Fig. 1-f, the gaps were counted out to a length of 31 words, and since the histogram would be symmetrical anyway, only the right half is plotted.

The expected height of the histogram bars, under the assumption of the statistical independence of the two words, is given by the middle horizontal line. The upper and lower horizontal lines represent deviations amounting to plus or minus the square root of the height of the middle line.

Discussion of the Data

It can be seen that, in general, the histograms show considerable deviation from what would be expected on the assumption of statistical independence. These deviations can be attributed to syntactic structure. Since our aim is to develop techniques that can be used in discovering structure, it is of interest to see how the deviations from randomness shown by the data correlate with what is known about the structure of English.

It two words occur together in a structure, that particular combination of two words will probably occur more frequently than expected on the assumption of statistical independence. The greater frequency of the particular combinations representing structures reduces the probability of occurrence of other combinations that do not represent structures.

Figures 1-a to 1-f all show a depressed region near the center of the histogram. This is taken to mean that these words tend not to recur immediately. The device of reduplication has only a limited use in English; this is probably true of many other languages. The length of the depressed region gives an idea of the length of the structures that frequently occur with these words. For example, structures with "the" can be expected to have two or three words. This is indeed true. But in the case of "and", the depressed region extends over at least 15 gaps. The total number of gaps of length 15 or less between occurrences of "and" amounts to only 50 as compared to an expected 68. This long depressed region can be understood as attributable to the fact that "and" not only correlates words, but longer structures as well. One of the uses of "and" is to coordinate clauses. Two occurrences of "and" used for this purpose cannot be closer together than the length of a clause.

Figure 1-a also shows that the word "the" has a slight periodicity with a gap length of 2 to 6. Such a periodicity can result from structures like the following:*

the tasks of the (2)
the linguist is to discover the (4)
the structure of languages or the (4)
the rules of the (2)
the difficulties that the (2)
the very complexity of the (3)

The average gap length between nearest occurrences of "the" is about 15 words. It is true that the depression at 0 and 1 must be compensated for by an increase elsewhere, but in the absence of other constraints, this increase would not cause a peak, but would be spread evenly over all other gap lengths. There would be fewer positions available for "the", but they would all be equally probable.

The gaps between different occurrences of the same word give an exactly symmetrical histogram, because it is always possible to interchange the words without altering their roles. Whenever two different words give an approximately symmetrical histogram, it gives us a clue that it may be possible to interchange them without altering their roles, i.e., they often play the same role and can be classed together.

In Figs. 1-g to 1-l, we have collected all the rest of the histograms that might be considered symmetrical. There is little question about the first four, but perhaps the last two do deviate from symmetry by more than the statistical fluctuations. Fig. 1-k shows possible deviations at gap lengths of one and two; Fig. 1-l shows possible deviations at a gap length of one. Let us assume that the top four are symmetrical. On this basis we tentatively group together and name:

---

* Examples are taken from the first paragraphs of this paper.

"the" and "a" (the article group)
"of" and "to" and "in" (the preposition group)

keeping "and" separate from all the rest on the assumption that Figs. 1-k and 1-l are not symmetrical.

All of the histograms of Fig. 2 are unsymmetrical. For two words to have an unsymmetrical gap histogram, they must frequently play different roles with respect to each other, and therefore they should not be grouped together.

Figures 2-a to 2-f are the six histograms that relate an article and a preposition. Our tentative grouping is given additional weight because these six histograms show certain similarities that can be attributed to the nature of articles and prepositions: The pattern "preposition article" occurs, and often with high frequency, while there are no cases of "article preposition".

These six histograms also show differences between the two articles and between the three prepositions. "The" is different from "a" in that it is the preferred article after "of". This shows up when Fig. 2-a is compared with Fig. 2-d. "Of" is different from "to", and from "in", which is probably a typical English preposition, in that "of" frequently follows an article with a gap of one or two. This is due to the very frequent "genitive" construction:

the tasks of
the structures of
the rules of
a grammar of

"To" is different from "of" and "in" in that it has a relatively low and broad peak before "the". The lowness of the peak is probably caused by the competition between the prepositional use of "to" and the use of "to" before an infinitive. The broadness is probably caused by an infinitive interposed between "to" and "the":

to discover the

Figures 2-g, 2-h, and 2-i show that "and" is different from "the", "a", and "in". If we take Fig. 1-k and Fig. 1-l as being unsymmetrical, it is also differentiated from "of" and "to".

The outlying peaks in Figs. 2-b and 2-e have not been explained. Perhaps they are statistical fluctuations. The accuracy of the counting has been verified.

Conclusions

The first trial of the use of gap analysis for revealing certain aspects of the syntax of a language has been quite fruitful in exposing certain ways in which the technique can be improved, and has been quite suggestive of its potentialities. The technique is certainly not a purely mechanical way of investigating the structure of language. A considerable amount of insight into language structure is required in order to make best use of the gap histograms as a tool of analysis.

The success of the procedure depends largely on the skill with which the text has been segmented into symbols. It is felt that this particular experiment would have been more meaningful if morphemes had been used instead of words as they are spelled. By using words, however, we eliminated much preliminary work. If one wants to use morphemes, perhaps it would be appropriate to segment into phonemes and use statistical procedures directly on the phonemes. The frequent morphemes would soon appear as frequent patterns of phonemes. If one sticks to conventional spelling, it would probably be better to include punctuation on a par with words.

One of the most serious limitations of our application of the procedure to English, was the shortness of the text that we chose. Conclusions could have been drawn with much greater certainty if statistical fluctuations had been smaller. Instead of 10,000 words, perhaps 100,000 words should be the minimum length of text for gap analysis. With a longer text, one could include many more of the frequent words because the word frequency distribution function begins to level off. Also with a longer text, one could take the next step of treating frequent constructions in the same manner as words and examining their effect on words in the vicinity. The construction "of the" and "the - of" were particularly frequent. They are probably as frequent as the 10th or 15th ranking word!

A systematic order of procedure for another experiment would be to count the frequencies of the words; then to count the gaps, taking first those that have the highest product of the frequencies of the words involved; then to look for frequent constructions and collate them into the list of word frequencies so that they could be used along with the words for further gap counts. By comparing the behavior of certain words in the vicinity of a two-word construction with the behavior of the words in the vicinity of each individual word involved in that construction, light can be shed on the multiple functions of words.

There are a number of other things that can be done with gap histograms. Certain features that many histograms or groups of histograms have in common, such as the "prepositional peak", can be used as a basis for grouping words together. Then the whole group of words can be counted as one to increase the number of cases, and rarer words can then be examined. On the other hand, the less frequent words can be grouped together on the basis of their behavior in the vicinity of frequent constructions. For example, one could group together all words that occur in the position between "the" and "of". Then the statistical behavior of the group could be investigated as if it were a single word. There is a certain resem-

blance here to the use of the substitution frame. Perhaps gap analysis will be able to reveal the best substitution frames for use with the more standard methods of linguistics.

Since the methods of gap analysis are far from being highly developed at this early stage, it is rather difficult to draw much of a comparison with the more standard linguistic methods involving informant techniques with native speakers. It is particularly difficult, if not impossible, to get accurate quantitative information from an informant. For this reason, grammars have made no pretense of being quantitative, but have merely reported what can occur and what cannot occur. Occasionally, linguists add the comment that something is "rare" or "usual" or a "favorite" construction. Statements like this reveal that they think that a certain amount of quantitative information is relevant, and that they would probably give more if they had the technique.

Gap analysis provides a wealth of numerical information, perhaps more than is really relevant. The linguist who uses it may have to pick and choose. Some types of numerical information about a text are of little significance. For example, the word "police" might be frequent in a newspaper, but the word "circuit" might be frequent in an electrical engineering article.

Because gap analysis is a numerical technique, it focuses attention on frequencies and numerical results. There is the continual implication that these numbers are worth something, that they are a relevant part of a grammar. To a certain extent this is true. In general, sentence structure is carried by combinations of the frequent morphemes. Infrequent words cannot indicate by their form alone their role in the sentence unless they have included in them some frequent role-marking morpheme. For example, a nonsense word like "sklack" could be a noun, a verb, an adjective, or an adverb. But "the sklack" or "sklacked", would have definite roles indicated by the frequent "the" or "ed". A sentence, then, can be considered as a structure of frequent morphemes with various open positions in it where all the rest of the morphemes can be put, including the infrequent and the new words. The frequent morphemes and their combinations can be considered as role markers for the less frequent ones. One can conclude that the frequent morphemes are the important ones for stating syntactic patterns, and that a careful use of gap analysis should be able to reveal these patterns.

References

1. Guiraud, P., "Bibliographie Critique de la Statistique Linguistique," Spectrum, Utrecht-Anvers, 1954.

2. Condon, E.U., "Statistics of Vocabulary," Science, 67, 300, (1928).

3. Mandelbrot, B., "Simple Games of Strategy Occurring in Communication through Natural Languages," Trans. I.R.E., PGIT-3, (March 1954), p.124.

4. Simon, H.A., "On a Class of Skew Distribution Functions," Biometrica, 42 (Dec. 1955), pp. 425-440.

5. Shannon, C.E., "Prediction and Entropy of Printed English," Bell Telephone System Monograph 1819 (1950).

6. Hockett, C.F., "A Manual of Phonology," Baltimore, Waverly Press, 1955, (Indiana University Publications in Anthropology and Linguistics, Memoir 11)

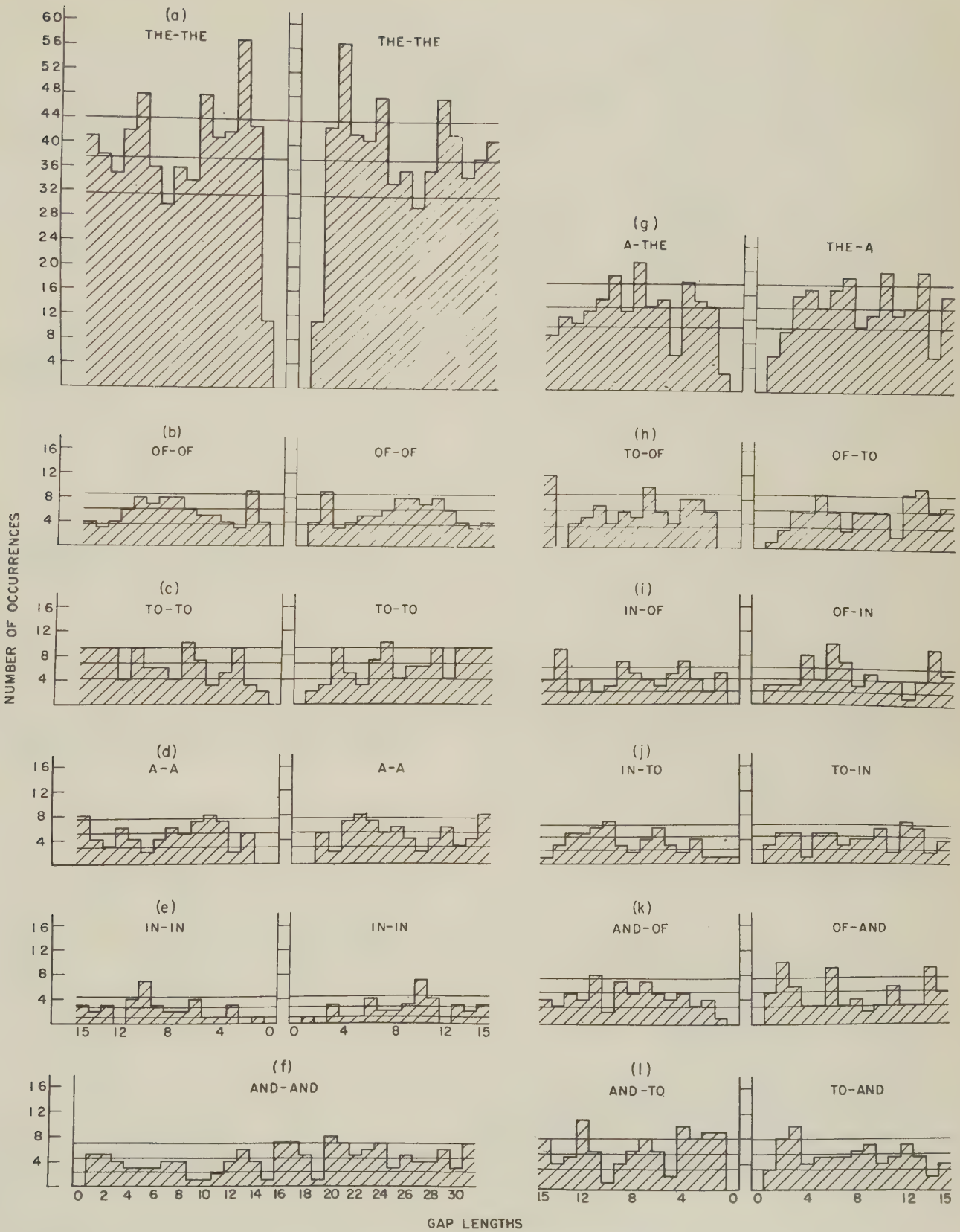


Fig. 1 - Gaps between occurrences of words.

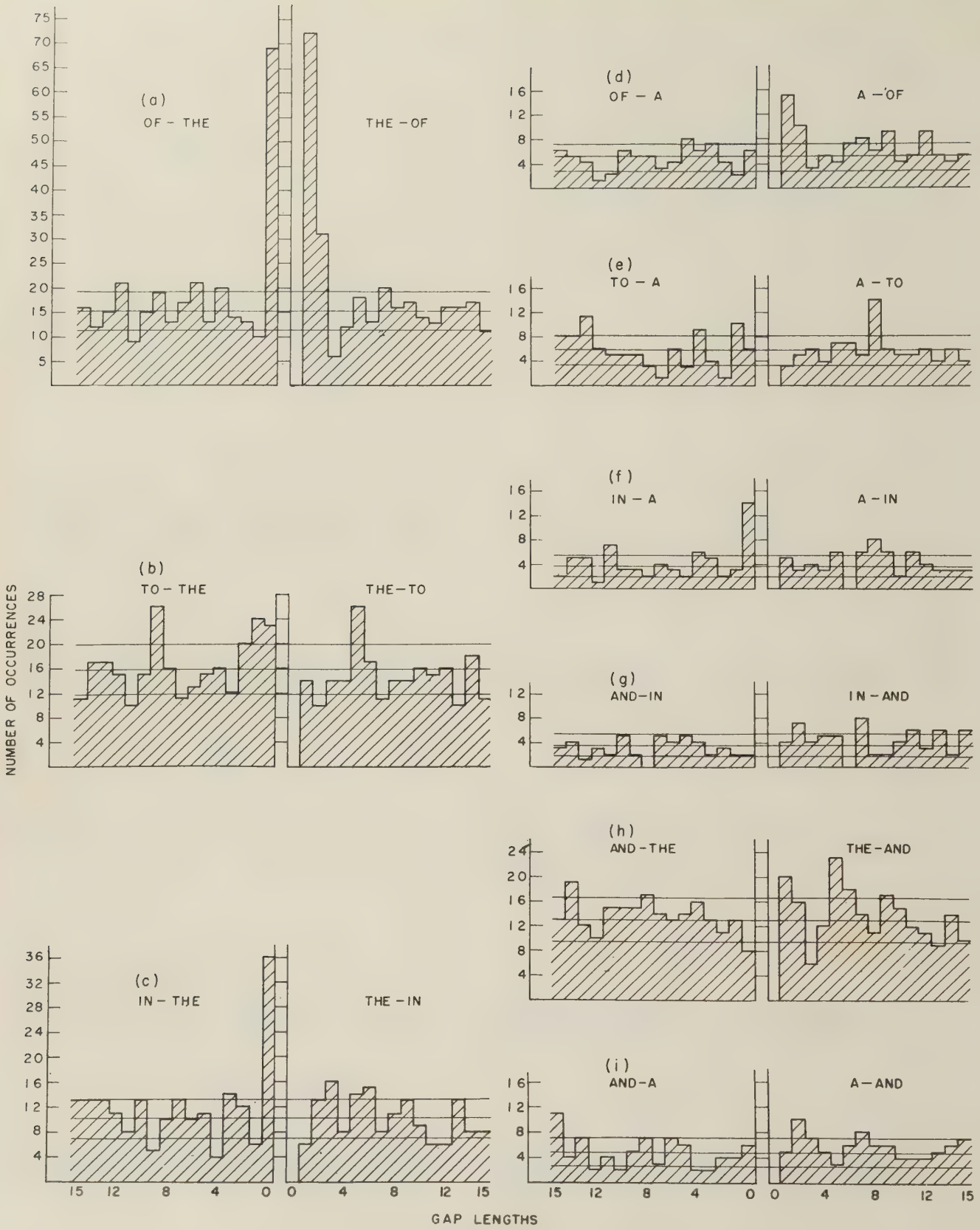


Fig. 2 - Gaps between occurrences of words.

THREE MODELS FOR THE DESCRIPTION OF LANGUAGE*

Noam Chomsky
Department of Modern Languages and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract

We investigate several conceptions of linguistic structure to determine whether or not they can provide simple and "revealing" grammars that generate all of the sentences of English and only these. We find that no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar. Furthermore, the particular subclass of such processes that produce n-order statistical approximations to English do not come closer, with increasing n, to matching the output of an English grammar. We formalize the notions of "phrase structure" and show that this gives us a method for describing language which is essentially more powerful, though still representable as a rather elementary type of finite-state process. Nevertheless, it is successful only when limited to a small subset of simple sentences. We study the formal properties of a set of grammatical transformations that carry sentences with phrase structure into new sentences with derived phrase structure, showing that transformational grammars are processes of the same elementary type as phrase-structure grammars; that the grammar of English is materially simplified if phrase structure description is limited to a kernel of simple sentences from which all other sentences are constructed by repeated transformations; and that this view of linguistic structure gives a certain insight into the use and understanding of language.

1. Introduction

There are two central problems in the descriptive study of language. One primary concern of the linguist is to discover simple and "revealing" grammars for natural languages. At the same time, by studying the properties of such successful grammars and clarifying the basic conceptions that underlie them, he hopes to arrive at a general theory of linguistic structure. We shall examine certain features of these related inquiries.

The grammar of a language can be viewed as a theory of the structure of this language. Any scientific theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations, to show how they are interrelated, and to predict an indefinite number of new phenomena. A mathematical theory has the additional property that predictions follow rigorously from the body of theory. Similarly, a grammar is based on a finite number of observed sentences (the linguist's corpus) and it "projects" this set to an infinite set of grammatical sentences by establishing general "laws" (grammatical rules) framed in terms of such hypothetical constructs as the particular phonemes, words, phrases, and so on, of the language under analysis. A properly formulated grammar should determine unambiguously the set of grammatical sentences.

General linguistic theory can be viewed as a metatheory which is concerned with the problem of how to choose such a grammar in the case of each particular language on the basis of a finite corpus of sentences. In particular, it will consider and attempt to explicate the relation between the set of grammatical sentences and the set of observed sentences. In other words, linguistic theory attempts to explain the ability of a speaker to produce and understand new sentences, and to reject as ungrammatical other new sequences, on the basis of his limited linguistic experience.

Suppose that for many languages there are certain clear cases of grammatical sentences and certain clear cases of ungrammatical sequences, e.g., (1) and (2), respectively, in English.

(1) John ate a sandwich
(2) Sandwich a ate John.

In this case, we can test the adequacy of a proposed linguistic theory by determining, for each language, whether or not the clear cases are handled properly by the grammars constructed in accordance with this theory. For example, if a large corpus of English does not happen to contain either (1) or (2), we ask whether the grammar that is determined for this corpus will project the corpus to include (1) and exclude (2). Even though such clear cases may provide only a weak test of adequacy for the grammar of a given language taken in isolation, they provide a very strong test for any general linguistic theory and for the set of grammars to which it leads, since we insist that in the case of each language the clear cases be handled properly in a fixed and predetermined manner. We can take certain steps towards the construction of an operational characterization of "grammatical sentence" that will provide us with the clear cases required to set the task of linguistics significantly.

*This work was supported in part by the Army (Signal Corps), the Air Force (Office of Scientific Research, Air Research and Development Command), and the Navy (Office of Naval Research), and in part by a grant from Eastman Kodak Company.

Observe, for example, that (1) will be read by an English speaker with the normal intonation of a sentence of the corpus, while (2) will be read with a falling intonation on each word, as will any sequence of unrelated words. Other distinguishing criteria of the same sort can be described.

Before we can hope to provide a satisfactory account of the general relation between observed sentences and grammatical sentences, we must learn a great deal more about the formal properties of each of these sets. This paper is concerned with the formal structure of the set of grammatical sentences. We shall limit ourselves to English, and shall assume intuitive knowledge of English sentences and nonsentences. We then ask what sort of linguistic theory is required as a basis for an English grammar that will describe the set of English sentences in an interesting and satisfactory manner.

The first step in the linguistic analysis of a language is to provide a finite system of representation for its sentences. We shall assume that this step has been carried out, and we shall deal with languages only in phonemic or alphabetic transcription. By a language, then, we shall mean a set (finite or infinite) of sentences, each of finite length, all constructed from a finite alphabet of symbols. If A is an alphabet, we shall say that anything formed by concatenating the symbols of A is a string in A. By a grammar of the language L we mean a device of some sort that produces all of the strings that are sentences of L and only these.

No matter how we ultimately decide to construct linguistic theory, we shall surely require that the grammar of any language must be finite. It follows that only a countable set of grammars is made available by any linguistic theory; hence that uncountably many languages, in our general sense, are literally not describable in terms of the conception of linguistic structure provided by any particular theory. Given a proposed theory of linguistic structure, then, it is always appropriate to ask the following question:

(3) Are there interesting languages that are simply outside the range of description of the proposed type?

In particular, we shall ask whether English is such a language. If it is, then the proposed conception of linguistic structure must be judged inadequate. If the answer to (3) is negative, we go on to ask such questions as the following:

(4) Can we construct reasonably simple grammars for all interesting languages?

(5) Are such grammars "revealing" in the sense that the syntactic structure that they exhibit can support semantic analysis, can provide insight into the use and understanding of language, etc.?

We shall first examine various conceptions of linguistic structure in terms of the possibility and complexity of description (questions (3), (4)). Then, in § 6, we shall briefly consider the same theories in terms of (5), and shall see that we are independently led to the same conclusions as to relative adequacy for the purposes of linguistics.

2. Finite State Markov Processes.

2.1 The most elementary grammars which, with a finite amount of apparatus, will generate an infinite number of sentences, are those based on a familiar conception of language as a particularly simple type of information source, namely, a finite-state Markov process.[1] Specifically, we define a finite-state grammar G as a system with a finite number of states $S_0,\dots,S_q$, a set $A=\{a_{ijk} \mid 0\leq i,j\leq q;\ 1\leq k\leq N_{ij}$ for each $i,j\}$ of transition symbols, and a set $C=\{(S_i,S_j)\}$ of certain pairs of states of G that are said to be connected. As the system moves from state S_i to S_j, it produces a symbol $a_{ijk}\in A$. Suppose that

(6) $S_{\alpha_1},\dots,S_{\alpha_m}$

is a sequence of states of G with $\alpha_1=\alpha_m=0$, and $(S_{\alpha_i},S_{\alpha_{i+1}})\in C$ for each $i<m$. As the system moves from S_{α_i} to $S_{\alpha_{i+1}}$ it produces the symbol

(7) $a_{\alpha_i\alpha_{i+1}k}$

for some $k\leq N_{\alpha_i\alpha_{i+1}}$. Using the arch $\frown$ to signify concatenation,[2] we say that the sequence (6) generates all sentences

(8) $a_{\alpha_1\alpha_2k_1}\frown a_{\alpha_2\alpha_3k_2}\frown\dots\frown a_{\alpha_{m-1}\alpha_mk_{m-1}}$

for all appropriate choices of k_i (i.e., for $k_i\leq N_{\alpha_i\alpha_{i+1}}$). The language L_G containing all and only such sentences is called the language generated by G.

Thus, to produce a sentence of L_G we set the system G in the initial state S_0 and we run through a sequence of connected states, ending again with S_0, and producing one of the associated transition symbols of A with each transition from one state to the next. We say that a language L is a finite-state language if L is the set of sentences generated by some finite-state grammar G.

2.2. Suppose that we take the set A of transition symbols to be the set of English phonemes. We can attempt to construct a finite state grammar G which will generate every string of English phonemes which is a grammatical sentence of English, and only such strings. It is immediately evident that the task of constructing a finite-state grammar for English can be considerably simplified if we take A as the set of English

morphemes[3] or words, and construct G so that it will generate exactly the grammatical strings of these units. We can then complete the grammar by giving a finite set of rules that give the phonemic spelling of each word or morpheme in each context in which it occurs. We shall consider briefly the status of such rules in § 4.1 and § 5.3.

Before inquiring directly into the problem of constructing a finite-state grammar for English morpheme or word sequences, let us investigate the absolute limits of the set of finite-state languages. Suppose that A is the alphabet of a language L, that $a_1,\ldots,a_n$ are symbols of this alphabet, and that $S=a_1^\frown\ldots^\frown a_n$ is a sentence of L. We say that S has an (i,j)-dependency with respect to L if and only if the following conditions are met:

(9)(i) $1 \leq i < j \leq n$

(ii) there are symbols $b_i, b_j \in A$ with the property that S_1 is not a sentence of L, and S_2 is a sentence of L, where S_1 is formed from S by replacing the i^{th} symbol of S (namely, a_i) by b_i, and S_2 is formed from S_1 by replacing the j^{th} symbol of S_1 (namely, a_j) by b_j.

In other words, S has an (i,j)-dependency with respect to L if replacement of the i^{th} symbol a_i of S by b_i ($b_i \neq a_i$) requires a corresponding replacement of the j^{th} symbol a_j of S by b_j ($b_j \neq a_j$) for the resulting string to belong to L.

We say that $D=\{(\alpha_1,\beta_1),\ldots,(\alpha_m,\beta_m)\}$ is a dependency set for S in L if and only if the following conditions are met:

(10)(i) For $1 \leq i \leq m$, S has an (α_i,β_i)-dependency with respect to L

(ii) for each i,j, $\alpha_i < \beta_j$

(iii) for each i,j such that $i \neq j$, $\alpha_i \neq \alpha_j$ and $\beta_i \neq \beta_j$.

Thus, in a dependency set for S in L every two dependencies are distinct in both terms and each "determining" element in S precedes all "determined" elements, where we picture a_{α_i} as determining the choice of a_{β_i}.

Evidently, if S has an m-termed dependency set in L, at least 2^m states are necessary in the finite-state grammar that generates the language L.

This observation enables us to state a necessary condition for finite-state languages.

(11) Suppose that L is a finite-state language. Then there is an m such that no sentence S of L has a dependency set of more than m terms in L.

With this condition in mind, we can easily construct many nonfinite-state languages. For example, the languages L_1, L_2, L_3 described in (12) are not describable by any finite-state grammar.

(12)(i) L_1 contains $a^\frown b$, $a^\frown a^\frown b^\frown b$, $a^\frown a^\frown a^\frown b^\frown b^\frown b$,..., and in general, all sentences consisting of n occurrences of a followed by exactly n occurrences of b, and only these;

(ii) L_2 contains $a^\frown a$, $b^\frown b$, $a^\frown b^\frown b^\frown a$, $b^\frown a^\frown a^\frown b$, $a^\frown a^\frown b^\frown b^\frown a^\frown a$,..., and in general, all "mirror-image" sentences consisting of a string X followed by X in reverse, and only these;

(iii) L_3 contains $a^\frown a$, $b^\frown b$, $a^\frown b^\frown a^\frown b$, $b^\frown a^\frown b^\frown a$, $a^\frown a^\frown b^\frown a^\frown a^\frown b$,..., and in general, all sentences consisting of a string X followed by the identical string X, and only these.

In L_2, for example, for any m we can find a sentence with a dependency set $D_m=\{(1,2m),(2,2m-1),\ldots,(m,m+1)\}$.[4]

2.3. Turning now to English, we find that there are infinite sets of sentences that have dependency sets with more than any fixed number of terms. For example, let $S_1,S_2,\ldots$ be declarative sentences. Then the following are all English sentences:

(13)(i) If S_1, then S_2.

(ii) Either S_3, or S_4.

(iii) The man who said that S_5, is arriving today.

These sentences have dependencies between "if"-"then", "either"-"or", "man"-"is". But we can choose S_1, S_3, S_5 which appear between the interdependent words, as (13i), (13ii), or (13iii) themselves. Proceeding to construct sentences in this way we arrive at subparts of English with just the mirror image properties of the languages L_1 and L_2 of (12). Consequently, English fails condition (11). English is not a finite-state language, and we are forced to reject the theory of language under discussion as failing condition (3).

We might avoid this consequence by an arbitrary decree that there is a finite upper limit to sentence length in English. This would serve no useful purpose, however. The point is that there are processes of sentence formation that this elementary model for language is intrinsically incapable of handling. If no finite limit is set for the operation of these processes, we can prove the literal inapplicability of this model. If the processes have a limit, then the construction of a finite-state grammar will not be literally impossible (since a list is a trivial finite-state grammar), but this grammar will be so complex as to be of little use or interest. Below, we shall study a model for grammars that can handle mirror-image languages. The extra power of such a model in the infinite case is reflected in the fact that it is much more useful and revealing if an upper limit is set. In general, the assumption that languages are infinite is made for the purpose of simplifying the description.[5] If a grammar has no recursive steps (closed loops, in the model

discussed above) it will be prohibitively complex--it will, in fact, turn out to be little better than a list of strings or of morpheme class sequences in the case of natural languages. If it does have recursive devices, it will produce infinitely many sentences.

2.4 Although we have found that no finite-state Markov process that produces sentences from left to right can serve as an English grammar, we might inquire into the possibility of constructing a sequence of such devices that, in some nontrivial way, come closer and closer to matching the output of a satisfactory English grammar. Suppose, for example, that for fixed n we construct a finite-state grammar in the following manner: one state of the grammar is associated with each sequence of English words of length n and the probability that the word X will be produced when the system is in the state S_i is equal to the conditional probability of X, given the sequence of n words which defines S_i. The output of such grammar is customarily called an $n+1^{st}$ order approximation to English. Evidently, as n increases, the output of such grammars will come to look more and more like English, since longer and longer sequences have a high probability of being taken directly from the sample of English in which the probabilities were determined. This fact has occasionally led to the suggestion that a theory of linguistic structure might be fashioned on such a model.

Whatever the other interest of statistical approximation in this sense may be, it is clear that it can shed no light on the problems of grammar. There is no general relation between the frequency of a string (or its component parts) and its grammaticalness. We can see this most clearly by considering such strings as

(14) colorless green ideas sleep furiously

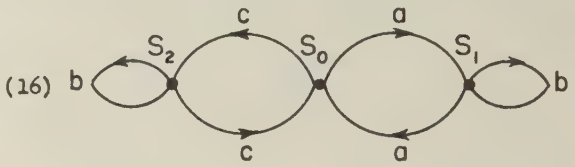
which is a grammatical sentence, even though it is fair to assume that no pair of its words may ever have occurred together in the past. Notice that a speaker of English will read (14) with the ordinary intonation pattern of an English sentence, while he will read the equally unfamiliar string

(15) furiously sleep ideas green colorless

with a falling intonation on each word, as in the case of any ungrammatical string. Thus (14) differs from (15) exactly as (1) differs from (2); our tentative operational criterion for grammaticalness supports our intuitive feeling that (14) is a grammatical sentence and that (15) is not. We might state the problem of grammar, in part, as that of explaining and reconstructing the ability of an English speaker to recognize (1), (14), etc., as grammatical, while rejecting (2), (15), etc. But no order of approximation model can distinguish (14) from (15) (or an indefinite number of similar pairs). As n increases, an n^{th} order approximation to English will exclude (as more and more improbable) an ever-increasing number of grammatical sentences, while it still contains vast numbers of completely ungrammatical strings.[6] We are forced to conclude that there is apparently no significant approach to the problems of grammar in this direction.

Notice that although for every n, a process of n-order approximation can be represented as a finite-state Markov process, the converse is not true. For example, consider the three-state process with (S_0,S_1), (S_1,S_1), (S_1,S_0), (S_0,S_2), (S_2,S_2), (S_2,S_0) as its only connected states, and with a, b, a, c, b, c as the respective transition symbols. This process can be represented by the following state diagram:

(16)



This process can produce the sentences a⌢a, a⌢b⌢a, a⌢b⌢b⌢a, a⌢b⌢b⌢b⌢a,..., c⌢c, c⌢b⌢c, c⌢b⌢b⌢c, c⌢b⌢b⌢b⌢c,..., but not a⌢b⌢b⌢c, c⌢b⌢b⌢a, etc. The generated language has sentences with dependencies of any finite length.

In § 2.4 we argued that there is no significant correlation between order of approximation and grammaticalness. If we order the strings of a given length in terms of order of approximation to English, we shall find both grammatical and ungrammatical strings scattered throughout the list, from top to bottom. Hence the notion of statistical approximation appears to be irrelevant to grammar. In § 2.3 we pointed out that a much broader class of processes, namely, all finite-state Markov processes that produce transition symbols, does not include an English grammar. That is, if we construct a finite-state grammar that produces only English sentences, we know that it will fail to produce an infinite number of these sentences; in particular, it will fail to produce an infinite number of true sentences, false sentences, reasonable questions that could be intelligibly asked, and the like. Below, we shall investigate a still broader class of processes that might provide us with an English grammar.

3. Phrase Structure.

3.1. Customarily, syntactic description is given in terms of what is called "immediate constituent analysis." In description of this sort the words of a sentence are grouped into phrases, these are grouped into smaller constituent phrases and so on, until the ultimate constituents (generally morphemes[3]) are reached. These phrases are then classified as noun phrases (NP), verb phrases (VP), etc. For example, the sentence (17) might be analyzed as in the accompanying diagram.

(17)

the man	took	the book
NP	Verb	NP
	VP	
Sentence		

Evidently, description of sentences in such terms permits considerable simplification over the word-by-word model, since the composition of a complex class of expressions such as NP can be stated just once in the grammar, and this class can be used as a building block at various points in the construction of sentences. We now ask what form of grammar corresponds to this conception of linguistic structure.

3.2. A phrase-structure grammar is defined by a finite vocabulary (alphabet) V_P, a finite set Σ of initial strings in V_P, and a finite set F of rules of the form: $X \rightarrow Y$, where X and Y are strings in V_P. Each such rule is interpreted as the instruction: rewrite X as Y. For reasons that will appear directly, we require that in each such $[\Sigma, F]$ grammar

(18) $\Sigma : \Sigma_1, \ldots, \Sigma_n$

$F: X_1 \rightarrow Y_1$
$\vdots$
$X_m \rightarrow Y_m$

Y_i is formed from X_i by the replacement of a single symbol of X_i by some string. Neither the replaced symbol nor the replacing string may be the identity element U of footnote 4.

Given the $[\Sigma, F]$ grammar (18), we say that:

(19)(i) a string β follows from a string α if $\alpha = Z^\frown X_i {}^\frown W$ and $\beta = Z^\frown Y_i {}^\frown W$, for some $i \leq m$;[7]

(ii) a derivation of the string S_t is a sequence $D=(S_1, \ldots, S_t)$ of strings, where $S_1 \in \Sigma$ and for each $i < t$, S_{i+1} follows from S_i;

(iii) a string S is derivable from (18) if there is a derivation of S in terms of (18);

(iv) a derivation of S_t is terminated if there is no string that follows from S_t;

(v) a string S_t is a terminal string if it is the last line of a terminated derivation.

A derivation is thus roughly analogous to a proof, with Σ taken as the axiom system and F as the rules of inference. We say that L is a derivable language if L is the set of strings that are derivable from some $[\Sigma, F]$ grammar, and we say that L is a terminal language if it is the set of terminal strings from some system $[\Sigma, F]$.

In every interesting case there will be a terminal vocabulary V_T ($V_T \subset V_P$) that exactly characterizes the terminal strings, in the sense that every terminal string is a string in V_T and no symbol of V_T is rewritten in any of the rules of F. In such a case we can interpret the terminal strings as constituting the language under analysis (with V_T as its vocabulary), and the derivations of these strings as providing their phrase structure.

3.3. As a simple example of a system of the form (18), consider the following small part of English grammar:

(20) Σ : #^Sentence^#
F: Sentence → NP^VP
VP → Verb^NP
NP → the^man, the^book
Verb → took

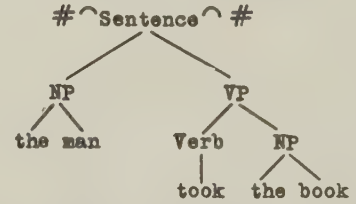
Among the derivations from (20) we have, in particular:

(21) D_1: #^Sentence^#
#^NP^VP^#
#^NP^Verb^NP^#
#^the^man^Verb^NP^#
#^the^man^Verb^the^book^#
#^the^man^took^the^book^#

D_2: #^Sentence^#
#^NP^VP^#
#^the^man^VP^#
#^the^man^Verb^NP^#
#^the^man^took^NP^#
#^the^man^took^the^book^#

These derivations are evidently equivalent; they differ only in the order in which the rules are applied. We can represent this equivalence graphically by constructing diagrams that correspond, in an obvious way, to derivations. Both D_1 and D_2 reduce to the diagram:

(22)



The diagram (22) gives the phrase structure of the terminal sentence "the man took the book," just as in (17). In general, given a derivation D of a string S, we say that a substring s of S is an X if in the diagram corresponding to D, s is traceable back to a single node, and this node is labelled X. Thus given D_1 or D_2, corresponding to (22), we say that "the^man" is an NP, "took^the^book" is a VP, "the^book" is an NP, "the^man^took^the^book" is a Sentence. "man^took," however, is not a phrase of this

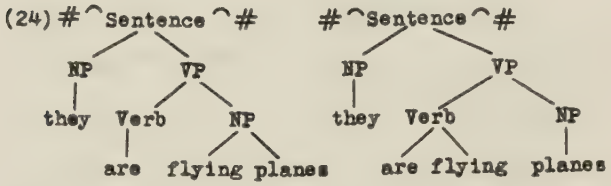
string at all, since it is not traceable back to any node.

When we attempt to construct the simplest possible [Σ ,F] grammar for English we find that certain sentences automatically receive non-equivalent derivations. Along with (20), the grammar of English will certainly have to contain such rules as

(23) Verb → are ^ flying
Verb → are
NP → they
NP → planes
NP → flying ^ planes

in order to account for such sentences as "they are flying - a plane" (NP-Verb-NP), "(flying) planes - are - noisy" (NP-Verb-Adjective), etc. But this set of rules provides us with two non-equivalent derivations of the sentence "they are flying planes", reducing to the diagrams:

(24)



Hence this sentence will have two phrase structures assigned to it; it can be analyzed as "they - are - flying planes" or "they - are flying - planes." And in fact, this sentence is ambiguous in just this way; we can understand it as meaning that "those specks on the horizon - are - flying planes" or "those pilots - are flying - planes." When the simplest grammar automatically provides nonequivalent derivations for some sentence, we say that we have a case of constructional homonymity, and we can suggest this formal property as an explanation for the semantic ambiguity of the sentence in question. In §1 we posed the requirement that grammars offer insight into the use and understanding of language (cf.(5)). One way to test the adequacy of a grammar is by determining whether or not the cases of constructional homonymity are actually cases of semantic ambiguity, as in (24) We return to this important problem in § 6.

In (20)-(24) the element # indicated sentence (later, word) boundary. It can be taken as an element of the terminal vocabulary V_T discussed in the final paragraph of § 3.2.

3.4. These segments of English grammar are much oversimplified in several respects. For one thing, each rule of (20) and (23) has only a single symbol on the left, although we placed no such limitation on [Σ ,F] grammars in § 3.2. A rule of the form

(25) $Z \hat{} X \hat{} W \rightarrow Z \hat{} Y \hat{} W$

indicates that X can be rewritten as Y only in the context Z--W. It can easily be shown that the grammar will be much simplified if we permit such rules. In § 3.2 we required that in such a rule as (25), X must be a single symbol. This ensures that a phrase-structure diagram will be constructible from any derivation. The grammar can also be simplified very greatly if we order the rules and require that they be applied in sequence (beginning again with the first rule after applying the final rule of the sequence), and if we distinguish between obligatory rules which must be applied when we reach them in the sequence and optional rules which may or may not be applied. These revisions do not modify the generative power of the grammar, although they lead to considerable simplification.

It seems reasonable to require for significance some guarantee that the grammar will actually generate a large number of sentences in a limited amount of time; more specifically, that it be impossible to run through the sequence of rules vacuously (applying no rule) unless the last line of the derivation under construction is a terminal string. We can meet this requirement by posing certain conditions on the occurrence of obligatory rules in the sequence of rules. We define a proper grammar as a system [Σ ,Q], where Σ is a set of initial strings and Q a sequence of rules $X_i \rightarrow Y_i$ as in (18), with the additional condition that for each i there must be at least one j such that $X_i = X_j$ and $X_j \rightarrow Y_j$ is an obligatory rule. Thus, each left-hand term of the rules of (18) must appear in at least one obligatory rule. This is the weakest simple condition that guarantees that a nonterminated derivation must advance at least one step every time we run through the rules. It provides that if X_i can be rewritten as one of $Y_{i_1}, \ldots, Y_{i_k}$ then at least one of these rewritings must take place. However, proper grammars are essentially different from [Σ ,F] grammars. Let D(G) be the set of derivations producible from a phrase structure grammar G, whether proper or not. Let $D_F = \{D(G) \mid G \text{ a } [\Sigma, F] \text{ grammar}\}$ and $D_Q = \{D(G) \mid G \text{ a proper grammar}\}$. Then

(26). D_F and D_Q are incomparable; i.e., $D_F \not\subset D_Q$ and $D_Q \not\subset D_F$.

That is, there are systems of phrase structure that can be described by [Σ ,F] grammars but not by proper grammars, and others that can be described by proper grammars but not by [Σ ,F] grammars.

3.5. We have defined three types of language: finite-state languages (in §2.1), derivable and terminal languages (in §3.2). These are related in the following way:

(27)(i) every finite-state language is a terminal language, but not conversely;
(ii) every derivable language is a terminal language, but not conversely;
(iii) there are derivable, nonfinite-state languages and finite-state, nonderivable languages.

Suppose that L_G is a finite-state language

with the finite-state grammar G as in § 2.1. We construct a $[\Sigma, F]$ grammar in the following manner: $\Sigma = \{S_0\}$; F contains a rule of the form (28i) for each i,j,k such that $(S_i, S_j) \in C$, $j \neq 0$, and $k \leq N_{ij}$; F contains a rule of the form (28ii) for each i,k such that $(S_i, S_0) \in C$ and $k \leq N_{i0}$.

$$(28)(i)\quad S_i \rightarrow a_{ijk} \frown S_j$$
$$(ii)\quad S_i \rightarrow a_{i0k}$$

Clearly, the terminal language from this $[\Sigma, F]$ grammar will be exactly L_G, establishing the first part of (27i).

In § 2.2 we found that L_1, L_2 and L_3 of (12) were not finite-state languages. L_1 and L_2, however, are terminal languages. For L_1, e.g., we have the $[\Sigma, F]$ grammar

$$(29)\quad \Sigma: Z \qquad F: Z \rightarrow a \frown b,\ Z \rightarrow a \frown Z \frown b$$

This establishes (27i).

Suppose that L_4 is a derivable language with the vocabulary $V_P = \{a_1, \ldots, a_n\}$. Suppose that we add to the grammar of L_4 a finite set of rules $a_i \rightarrow b_i$, where the b_i's are not in V_P and are all distinct. Then this new grammar gives a terminal language which is simply a notational variant of L_4. Thus every derivable language is also terminal.

As an example of a terminal, nonderivable language consider the language L_5 containing just the strings

$$(30)\quad a \frown b,\ c \frown a \frown b \frown d,\ c \frown c \frown a \frown b \frown d \frown d,\ c \frown c \frown c \frown a \frown b \frown d \frown d \frown d, \ldots$$

An infinite derivable language must contain an infinite set of strings that can be arranged in a sequence $S_1, S_2, \ldots$ in such a way that for some rule $X \rightarrow Y$, S_i follows from S_{i-1} by application of this rule, for each $i > 1$. And Y in this rule must be formed from X by replacement of a <u>single symbol</u> of X by a string (cf. (18)). This is evidently impossible in the case of L_5. This language is, however, the terminal language given by the following grammar:

$$(31)\quad \Sigma: Z \qquad F: Z \rightarrow a \frown b,\ Z \rightarrow c \frown Z \frown d$$

An example of a finite-state, nonderivable language is the language L_6 containing all and only the strings consisting of 2n or 3n occurrences of a, for n=1,2,.... Language L_1 of (12) is a derivable, nonfinite-state language, with the initial string $a \frown b$ and the rule: $a \frown b \rightarrow a \frown a \frown b \frown b$.

The major import of Theorem (27) is that description in terms of phrase structure is essentially more powerful (not just simpler) than description in terms of the finite-state grammars that produce sentences from left to right. In § 2.3 we found that English is literally beyond the bounds of these grammars because of mirror-image properties that it shares with L_1 and L_2 of (12). We have just seen, however, that L_1 is a terminal language and the same is true of L_2. Hence, the considerations that led us to reject the finite-state model do not similarly lead us to reject the more powerful phrase-structure model.

Note that the latter is more abstract than the finite-state model in the sense that symbols that are not included in the vocabulary of a language enter into the description of this language. In the terms of § 3.2, V_P properly includes V_T. Thus in the case of (29), we describe L_1 in terms of an element Z which is not in L_1; and in the case of (20)-(24), we introduce such symbols as Sentence, NP, VP, etc., which are not words of English, into the description of English structure.

3.6. We can interpret a $[\Sigma, F]$ grammar of the form (18) as a rather elementary finite-state process in the following way. Consider a system that has a finite number of states $S_0, \ldots, S_q$. When in state S_0, it can produce any of the strings of Σ, thereby moving into a new state. Its state at any point is determined by the subset of elements of $X_1, \ldots, X_m$ contained as substrings in the last produced string, and it moves to a new state by applying one of the rules to this string, thus producing a new string. The system returns to state S_0 with the production of a terminal string. This system thus produces derivations, in the sense of § 3.2. The process is determined at any point by its present state and by the last string that has been produced, and there is a finite upper bound on the amount of inspection of this string that is necessary before the process can continue, producing a new string that differs in one of a finite number of ways from its last output.

It is not difficult to construct languages that are beyond the range of description of $[\Sigma, F]$ grammars. In fact, the language L_3 of (12iii) is evidently not a terminal language. I do not know whether English is actually a terminal language or whether there are other actual languages that are literally beyond the bounds of phrase structure description. Hence I see no way to disqualify this theory of linguistic structure on the basis of consideration (3). When we turn to the question of the complexity of description (cf. (4)), however, we find that there are ample grounds for the conclusion that this theory of linguistic structure is fundamentally inadequate. We shall now investigate a few of the problems

that arise when we attempt to extend (20) to a full-scale grammar of English.

4. Inadequacies of Phrase-Structure Grammar

4.1. In (20) we considered only one way of developing the element Verb, namely, as "took". But even with the verb stem fixed there are a great many other forms that could appear in the context "the man -- the book," e.g., "takes," "has taken," "has been taking," "is taking," "has been taken," "will be taking," and so on. A direct description of this set of elements would be fairly complex, because of the heavy dependencies among them (e.g., "has taken" but not "has taking," "is being taken" but not "is being taking," etc.). We can, in fact, give a very simple analysis of "Verb" as a sequence of independent elements, but only by selecting as elements certain discontinuous strings. For example, in the phrase "has been taking" we can separate out the discontinuous elements "has..en," "be..ing," and "take", and we can then say that these elements combine freely. Following this course systematically, we replace the last rule in (20) by

(32) (i) Verb → Auxiliary ^ V
(ii) V → take, eat,...
(iii) Auxiliary → C(M)(have ^ en) (be ^ ing) (be ^ en)
(iv) M → will, can, shall, may, must
(v) C → past, present

The notations in (32iii) are to be interpreted as follows: in developing "Auxiliary" in a derivation we must choose the unparenthesized element C, and we may choose zero or more of the parenthesized elements, in the order given. Thus, in continuing the derivation D_1 of (21) below line five, we might proceed as follows:

(33) # ^ the ^ man ^ Verb ^ the ^ book ^ #
[from D_1 of (21)]
^ the ^ man ^ Auxiliary ^ V ^ the ^ book ^
[(32i)]
^ the ^ man ^ Auxiliary ^ take ^ the ^ book ^
[(32ii)]
^ the ^ man ^ C ^ have ^ en ^ be ^ ing ^ take ^ the ^ book ^
[(32iii), choosing the elements C, have ^ en, and be ^ ing]
^ the ^ man ^ past ^ have ^ en ^ be ^ ing ^ take ^ the ^ book ^
[(32v)]

Suppose that we define the class Af as containing the affixes "en", "ing", and the C's; and the class v as including all V's, M's, "have", and "be." We can then convert the last line of (33) into a properly ordered sequence of morphemes by the following rule:

(34) Af ^ v → v ^ Af ^ #

Applying this rule to each of the three Af ^ v sequences in the last line of (33), we derive

(35) # ^ the ^ man ^ have ^ past ^ # ^ be ^ en ^ # ^ take ^ ing ^ # ^ the ^ book ^ #.

In the first paragraph of § 2.2 we mentioned that a grammar will contain a set of rules (called morphophonemic rules) which convert strings of morphemes into strings of phonemes. In the morphophonemics of English, we shall have such rules as the following (we use conventional, rather than phonemic orthography):

(36) have ^ past → had
be ^ en → been
take ^ ing → taking
will ^ past → would
can ^ past → could
M ^ present → M
walk ^ past → walked
take ^ past → took
etc.

Applying the morphophonemic rules to (35) we derive the sentence:

(37) the man had been taking the book.

Similarly, with one major exception to be discussed below (and several minor ones that we shall overlook here), the rules (32), (34) will give all the other forms of the verb in declarative sentences, and only these forms.

This very simple analysis, however, goes beyond the bounds of [Σ ,F] grammars in several respects. The rule (34), although it is quite simple, cannot be incorporated within a [Σ ,F] grammar, which has no place for discontinuous elements. Furthermore, to apply the rule (34) to the last line of (33) we must know that "take" is a V, hence, a v. In other words, in order to apply this rule it is necessary to inspect more than just the string to which the rule applies; it is necessary to know some of the constituent structure of this string, or equivalently (cf. § 3.3), to inspect certain earlier lines in its derivation. Since (34) requires knowledge of the 'history of derivation' of a string, it violates the elementary property of [Σ ,F] grammars discussed in § 3.6.

4.2. The fact that this simple analysis of the verb phrase as a sequence of independently chosen units goes beyond the bounds of [Σ ,F] grammars, suggests that such grammars are too limited to give a true picture of linguistic structure. Further study of the verb phrase lends additional support to this conclusion. There is one major limitation on the independence of the elements introduced in (32). If we choose an intransitive verb (e.g., "come," "occur," etc.) as V in (32), we cannot select be ^ en as an auxiliary. We cannot have such phrases as "John has been come," "John is occurred," and the like. Furthermore, the element be ^ en cannot be chosen independently of the context of the phrase "Verb." If we have

the element "Verb" in the context "the man -- the food," we are constrained not to select be⌢en in applying (32), although we are free to choose any other element of (32). That is, we can have "the man is eating the food," "the man would have been eating the food," etc., but not "the man is eaten the food," "the man would have been eaten the food," etc. On the other hand, if the context of the phrase "Verb" is, e.g., "the food -- by the man," we are required to select be⌢en. We can have "the food is eaten by the man," but not "the food is eating by the man," etc. In short, we find that the element be⌢en enters into a detailed network of restrictions which distinguish it from all the other elements introduced in the analysis of "Verb" in (32). This complex and unique behavior of be⌢en suggests that it would be desirable to exclude it from (32) and to introduce passives into the grammar in some other way.

There is, in fact, a very simple way to incorporate sentences with be⌢en (i.e., passives) into the grammar. Notice that for every active sentence such as "the man ate the food" we have a corresponding passive "the food was eaten by the man" and conversely. Suppose then that we drop the element be⌢en from (32iii), and then add to the grammar the following rule:

(38) If S is a sentence of the form NP_1-Auxiliary-V-NP_2, then the corresponding string of the form NP_2-Auxiliary⌢be⌢en-V-by⌢NP_1 is also a sentence.

For example, if "the man - past - eat the food" (NP_1-Auxiliary-V-NP_2) is a sentence, then "the food - past be en - eat - by the man" (NP_2-Auxiliary⌢be⌢en-V-by⌢NP_1) is also a sentence. Rules (34) and (36) would convert the first of these into "the man ate the food" and the second into "the food was eaten by the man."

The advantages of this analysis of passives are unmistakable. Since the element be⌢en has been dropped from (32) it is no longer necessary to qualify (32) with the complex of restrictions discussed above. The fact that be⌢en can occur only with transitive verbs, that it is excluded in the context "the man -- the food" and that it is required in the context "the food -- by the man," is now, in each case, an automatic consequence of the analysis we have just given.

A rule of the form (38), however, is well beyond the limits of phrase-structure grammars. Like (34), it rearranges the elements of the string to which it applies, and it requires considerable information about the constituent structure of this string. When we carry the detailed study of English syntax further, we find that there are many other cases in which the grammar can be simplified if the [Σ ,F] system is supplemented by rules of the same general form as (38). Let us call each such rule a grammatical transformation. As our third model for the description of linguistic structure, we now consider briefly the formal properties of a transformational grammar that can be adjoined to the [Σ ,F] grammar of phrase structure.[8]

5. Transformational Grammar.

5.1. Each grammatical transformation T will essentially be a rule that converts every sentence with a given constituent structure into a new sentence with derived constituent structure. The transform and its derived structure must be related in a fixed and constant way to the structure of the transformed string, for each T. We can characterize T by stating, in structural terms, the domain of strings to which it applies and the change that it effects on any such string.

Let us suppose in the following discussion that we have a [Σ ,F] grammar with a vocabulary V_P and a terminal vocabulary $V_T \subset V_P$, as in § 3.2.

In § 3.3 we showed that a [Σ ,F] grammar permits the derivation of terminal strings, and we pointed out that in general a given terminal string will have several equivalent derivations. Two derivations were said to be equivalent if they reduce to the same diagram of the form (22), etc.[9] Suppose that $D_1,..,D_n$ constitute a maximal set of equivalent derivations of a terminal string S. Then we define a phrase marker of S as the set of strings that occur as lines in the derivations $D_1,...,D_n$. A string will have more than one phrase marker if and only if it has nonequivalent derivations (cf. (24)).

Suppose that K is a phrase marker of S. We say that

(39) (S,K) is analyzable into $(X_1,...,X_n)$ if and only if there are strings $s_1,..,s_n$ such that
(i) $S=s_1 \frown ... \frown s_n$
(ii) for each $i \leq n$, K contains the string
$$s_1 \frown .. \frown s_{i-1} \frown X_i \frown s_{i+1} \frown .. \frown s_n$$

(40) In this case, s_i is an X_i in S with respect to K.[10]

The relation defined in (40) is exactly the relation "is a" defined in § 3.3; i.e., s_i is an X_i in the sense of (40) if and only if s_i is a substring of S which is traceable back to a single node of the diagram of the form (22), etc., and this node is labelled X_i.

The notion of analyzability defined above allows us to specify precisely the domain of application of any transformation. We associate with each transformation a restricting class R defined as follows:

(41) R is a restricting class if and only if for some r,m, R is the set of sequences:

$$X_1^1,...,X_r^1$$
$$\vdots$$
$$X_1^m,...,X_r^m$$

where X_i^j is a string in the vocabulary V_P, for each i,j. We then say that a string S with the phrase marker K belongs to the domain of the transformation

T if the restricting class R associated with T contains a sequence $(X_1^j,\ldots,X_r^j)$ into which (S,K) is analyzable. The domain of a transformation is thus a set of ordered pairs (S,K) of a string S and a phrase marker K of S. A transformation may be applicable to S with one phrase marker, but not with a second phrase marker, in the case of a string S with ambiguous constituent structure.

In particular, the passive transformation described in (38) has associated with it a restricting class R_p containing just one sequence:

(42) $R_p = \left\{ (\text{NP, Auxiliary, V, NP}) \right\}$.

This transformation can thus be applied to any string that is analyzable into an NP followed by an Auxiliary followed by a V followed by an NP. For example, it can be applied to the string (43) analyzed into substrings $s_1,\ldots,s_4$ in accordance with the dashes.

(43) the man - past - eat - the food.

5.2. In this way, we can describe in structural terms the set of strings (with phrase markers) to which any transformation applies. We must now specify the structural change that a transformation effects on any string in its domain. An <u>elementary transformation</u> t is defined by the following property:

(44) for each pair of integers n,r ($n \leq r$), there is a unique sequence of integers $(a_0,a_1,\ldots,a_k)$ and a unique sequence of strings in V_p $(Z_1,\ldots,Z_{k+1})$ such that (i)$a_0=0$; $k \geq 0$; $1 \leq a_j \leq r$ for $1 \leq j \leq k$; $Y_0 = U$[11]

(ii) for each $Y_1,\ldots,Y_r$,

$t(Y_1,\ldots,Y_n;Y_n,\ldots,Y_r) = Y_{a_0}^\frown Z_1^\frown Y_{a_1}^\frown Z_2^\frown Y_{a_2}^\frown \ldots^\frown Y_{a_k}^\frown Z_{k+1}$:

Thus t can be understood as converting the occurrence of Y_n in the context

(45) $Y_1^\frown \ldots^\frown Y_{n-1}^\frown -\!-^\frown Y_{n+1}^\frown \ldots^\frown Y_r$

into a certain string $Y_{a_0}^\frown Z_1^\frown \ldots^\frown Y_{a_k}^\frown Z_{k+1}$

which is unique, given the sequence of terms $(Y_1,\ldots,Y_r)$ into which $Y_1^\frown \ldots^\frown Y_r$ is subdivided. t carries the string $Y_1-\ldots-Y_r$ into a new string $W_1-\ldots-W_r$ which is related in a fixed way to $Y_1-\ldots-Y_r$. More precisely, we associate with t the derived transformation t*:

(46) t* is the <u>derived transformation</u> of t if and only if for all $Y_1,\ldots,Y_r$, $t^*(Y_1,\ldots,Y_r) = W_1^\frown \ldots^\frown W_r$, where $W_n = t(Y_1,\ldots,Y_n;Y_n,\ldots,Y_r)$ for each $n \leq r$.

We now associate with each transformation T an elementary transformation t. For example, with the passive transformation (38) we associate the elementary transformation t_p defined as follows:

(47) $t_p(Y_1;Y_1,\ldots,Y_4) = Y_4$

$t_p(Y_1,Y_2;Y_2,Y_3,Y_4) = Y_2^\frown be^\frown en$

$t_p(Y_1,Y_2,Y_3;Y_3,Y_4) = Y_3$

$t_p(Y_1,\ldots,Y_4;Y_4) = by^\frown Y_1$

$t_p(Y_1,\ldots,Y_n;Y_n,\ldots,Y_r) = Y_n$ for all $n \leq r \neq 4$.

The derived transformation t_p^* thus has the following effect:

(48)(i) $t_p^*(Y_1,\ldots,Y_4) = Y_1 - Y_2^\frown be^\frown en - Y_3 - by^\frown Y_1$

(ii) $t_p^*(\text{the}^\frown \text{man, past, eat, the}^\frown \text{food}) =$ $\text{the}^\frown \text{food} - \text{past}^\frown be^\frown en - \text{eat} - by^\frown \text{the}^\frown \text{man}$.

The rules (34),(36) carry the right-hand side of (48ii) into "the food was eaten by the man," just as they carry (43) into the corresponding active "the man ate the food."

The pair (R_p,t_p) as in (42),(47) completely characterizes the passive transformation as described in (38). R_p tells us to which strings this transformation applies (given the phrase markers of these strings) and how to subdivide these strings in order to apply the transformation, and t_p tells us what structural change to effect on the subdivided string.

A grammatical transformation is specified completely by a restricting class R and an elementary transformation t, each of which is finitely characterizable, as in the case of the passive. It is not difficult to define rigorously the manner of this specification, along the lines sketched above. To complete the development of transformational grammar it is necessary to show how a transformation automatically assigns a derived phrase marker to each transform and to generalize to transformations on sets of strings. (These and related topics are treated in reference [3].) A transformation will then carry a string S with a phrase marker K (or a set of such pairs) into a string S' with a derived phrase marker K'.

5.3. From these considerations we are led to a picture of grammars as possessing a tripartite structure. Corresponding to the phrase structure analysis we have a sequence of rules of the form X→Y, e.g., (20), (23), (32). Following this we have a sequence of transformational rules such as (34) and (38). Finally, we have a sequence of morphophonemic rules such as (36), again of the form X→Y. To generate a sentence from such a grammar we construct an extended derivation beginning with an initial string of the phrase structure grammar, e.g., #^Sentence^#, as in (20). We then run through the rules of phrase structure, producing a terminal string. We then apply certain transformations, giving a string of morphemes in the correct order, perhaps quite a different string from the original terminal string. Application of the morphophonemic rules converts this into a string of phonemes. We might run through the phrase structure grammar several times and then apply a generalized transformation to the resulting set of terminal strings.

In § 3.4 we noted that it is advantageous to order the rules of phrase structure into a sequence, and to distinguish obligatory from optional rules. The same is true of the transformational part of the grammar. In § 4 we discussed the transformation (34), which converts a

sequence affix-verb into the sequence verb-affix, and the passive transformation (38). Notice that (34) must be applied in every extended derivation, or the result will not be a grammatical sentence. Rule (34), then, is an obligatory transformation. The passive transformation, however, may or may not be applied; either way we have a sentence. The passive is thus an optional transformation. This distinction between optional and obligatory transformations leads us to distinguish between two classes of sentences of the language. We have, on the one hand, a kernel of basic sentences that are derived from the terminal strings of the phrase-structure grammar by application of only obligatory transformations. We then have a set of derived sentences that are generated by applying optional transformations to the strings underlying kernel sentences.

When we actually carry out a detailed study of English structure, we find that the grammar can be greatly simplified if we limit the kernel to a very small set of simple, active, declarative sentences (in fact, probably a finite set) such as "the man ate the food," etc. We then derive questions, passives, sentences with conjunction, sentences with compound noun phrases (e.g., "proving that theorem was difficult," with the NP "proving that theorem"),[12] etc., by transformations. Since the result of a transformation is a sentence with derived constituent structure, transformations can be compounded, and we can form questions from passives (e.g., "was the food eaten by the man"), etc. The actual sentences of real life are usually not kernel sentences, but rather complicated transforms of these. We find, however, that the transformations are, by and large, meaning-preserving, so that we can view the kernel sentences underlying a given sentence as being, in some sense, the elementary "content elements" in terms of which the actual transform is "understood." We discuss this problem briefly in § 6, more extensively in references [1], [2].

In § 3.6 we pointed out that a grammar of phrase structure is a rather elementary type of finite-state process that is determined at each point by its present state and a bounded amount of its last output. We discovered in § 4 that this limitation is too severe, and that the grammar can be simplified by adding transformational rules that take into account a certain amount of constituent structure (i.e., a certain history of derivation). However, each transformation is still finitely characterizable (cf. §§ 5.1-2), and the finite restricting class (41) associated with a transformation indicates how much information about a string is needed in order to apply this transformation. The grammar can therefore still be regarded as an elementary finite-state process of the type corresponding to phrase structure. There is still a bound, for each grammar, on how much of the past output must be inspected in order for the process of derivation to continue, even though more than just the last output (the last line of the derivation) must be known.

6. Explanatory Power of Linguistic Theories

We have thus far considered the relative adequacy of theories of linguistic structure only in terms of such essentially formal criteria as simplicity. In § 1 we suggested that there are other relevant considerations of adequacy for such theories. We can ask (cf.(5)) whether or not the syntactic structure revealed by these theories provides insight into the use and understanding of language. We can barely touch on this problem here, but even this brief discussion will suggest that this criterion provides the same order of relative adequacy for the three models we have considered.

If the grammar of a language is to provide insight into the way the language is understood, it must be true, in particular, that if a sentence is ambiguous (understood in more than one way), then this sentence is provided with alternative analyses by the grammar. In other words, if a certain sentence S is ambiguous, we can test the adequacy of a given linguistic theory by asking whether or not the simplest grammar constructible in terms of this theory for the language in question automatically provides distinct ways of generating the sentence S. It is instructive to compare the Markov process, phrase-structure, and transformational models in the light of this test.

In § 3.3 we pointed out that the simplest $[\Sigma, F]$ grammar for English happens to provide nonequivalent derivations for the sentence "they are flying planes," which is, in fact, ambiguous. This reasoning does not appear to carry over for finite-state grammars, however. That is, there is no obvious motivation for assigning two different paths to this ambiguous sentence in any finite-state grammar that might be proposed for a part of English. Such examples of constructional homonymity (there are many others) constitute independent evidence for the superiority of the phrase-structure model over finite-state grammars.

Further investigation of English brings to light examples that are not easily explained in terms of phrase structure. Consider the phrase

(49) the shooting of the hunters.

We can understand this phrase with "hunters" as the subject, analogously to (50), or as the object, analogously to (51).

(50) the growling of lions

(51) the raising of flowers.

Phrases (50) and (51), however, are not similarly ambiguous. Yet in terms of phrase structure, each of these phrases is represented as: the - V⌢ing - of⌢NP.

Careful analysis of English shows that we can simplify the grammar if we strike the phrases (49)-(51) out of the kernel and reintroduce them transformationally by a transformation T_1 that carries such sentences as "lions growl" into (50), and a transformation T_2 that carries such sentences

as "they raise flowers" into (51). T_1 and T_2 will be similar to the nominalizing transformation described in fn.12, when they are correctly constructed. But both "hunters shoot" and "they shoot the hunters" are kernel sentences; and application of T_1 to the former and T_2 to the latter yields the result (49). Hence (49) has two distinct transformational origins. It is a case of constructional homonymity on the transformational level. The ambiguity of the grammatical relation in (49) is a consequence of the fact that the relation of "shoot" to "hunters" differs in the two underlying kernel sentences. We do not have this ambiguity in the case of (50), (51), since neither "they growl lions" nor "flowers raise" is a grammatical kernel sentence.

There are many other examples of the same general kind (cf. [1],[2]), and to my mind, they provide quite convincing evidence not only for the greater adequacy of the transformational conception of linguistic structure, but also for the view expressed in §5.4 that transformational analysis enables us to reduce partially the problem of explaining how we understand a sentence to that of explaining how we understand a kernel sentence.

In summary, then, we picture a language as having a small, possibly finite kernel of basic sentences with phrase structure in the sense of §3, along with a set of transformations which can be applied to kernel sentences or to earlier transforms to produce new and more complicated sentences from elementary components. We have seen certain indications that this approach may enable us to reduce the immense complexity of actual language to manageable proportions and, in addition, that it may provide considerable insight into the actual use and understanding of language.

Footnotes

1. Cf. [7]. Finite-state grammars can be represented graphically by state diagrams, as in [7], p.15f.

2. See [6], Appendix 2, for an axiomatization of concatenation algebras.

3. By 'morphemes' we refer to the smallest grammatically functioning elements of the language, e.g., "boy", "run", "ing" in "running", "s" in "books", etc.

4. In the case of L_1, b_j of (9ii) can be taken as an identity element U which has the property that for all X, $U^\frown X = X^\frown U = X$. Then D_m will also be a dependency set for a sentence of length 2m in L_1.

5. Note that a grammar must reflect and explain the ability of a speaker to produce and understand new sentences which may be much longer than any he has previously heard.

6. Thus we can always find sequences of n+1 words whose first n words and last n words may occur, but not in the same sentence (e.g. replace "is" by "are" in (13iii), and choose S_5 of any required length).

7. Z or W may be the identity element U (cf. fn.4) in this case. Note that since we limited (18) so as to exclude U from figuring significantly on either the right- or the left-hand side of a rule of F, and since we required that only a single symbol of the left-hand side may be replaced in any rule, it follows that Y_i must be at least as long as X_i. Thus we have a simple decision procedure for derivability and terminality in the sense of (19iii), (19v).

8. See [3] for a detailed development of an algebra of transformations for linguistic description and an account of transformational grammar. For further application of this type of description to linguistic material, see [1], [2], and from a somewhat different point of view, [4].

9. It is not difficult to give a rigorous definition of the equivalence relation in question, though this is fairly tedious.

10. The notion "is a" should actually be relativized further to a given occurrence of s_i in S. We can define an occurrence of s_i in S as an ordered pair (s_i, X), where X is an initial substring of S, and s_i is a final substring of X. Cf. [5], p.297.

11. Where U is the identity, as in fn. 4.

12. Notice that this sentence requires a generalized transformation that operates on a pair of strings with their phrase markers. Thus we have a transformation that converts S_1, S_2 of the forms NP-VP_1, it-VP_2, respectively, into the string: ing$^\frown VP_1$ - VP_2. It converts S_1 = "they - prove that theorem", S_2 = "it - was difficult" into "ing prove that theorem - was difficult," which by (34) becomes "proving that theorem was difficult." Cf. [1], [3] for details.

Bibliography

[1] Chomsky, N., The Logical Structure of Linguistic Theory (mimeographed).

[2] Chomsky, N., Syntactic Structures, to be published by Mouton & Co., 'S-Gravenhage, Netherlands.

[3] Chomsky, N., Transformational Analysis, Ph. D. Dissertation, University of Pennsylvania, June, 1955.

[4] Harris, Z.S., Discourse Analysis, Language 28,1 (1952).

[5] Quine, W.V., Mathematical Logic, revised edition, Harvard University Press., Cambridge, 1951.

[6] Rosenbloom, P., Elements of Mathematical Logic, Dover, New York, 1950

[7] Shannon & Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, 1949.

SOME STUDIES IN THE SPEED OF VISUAL PERCEPTION

George C. Sziklai
RCA Laboratories Division
Princeton, N.J.

Summary

Statistical studies of television signals indicated a high degree of correlation between successive elements, lines and frames. A continuous test run indicated that the normalized detail content[1] of actual broadcast signals run lower than 5 percent.

The fact that most transmitted television scenes thus measured contained very little detail compared to some artificial subjects indicated that the observer's preference has influenced the subject to be transmitted. In order to verify this assumption, some tests were devised to measure the perception speed of observers. These tests included certain reading and character recognition tests and finally a test consisting of object recognition in precisely measured periods was devised.

In order to evaluate the perception rate in bits, a relationship between the gestalt concept and the binary choices was sought. Assuming a limited number of picturable nouns to correspond to the gestalt experience of the observer, each object recognized was assigned a value of ten bits (corresponding to 1024 picturable nouns). Several series of these tests indicated that the visual perception speed of a normal observer is between 30 to 50 bits per second, that this value holds for periods of one-tenth to two seconds, and that the first thing observed is the center of the picture. These values are similar to the values obtained by Licklider, Pierce and others for reading speeds.

Introduction

The statement that the television signal is highly redundant and therefore very wasteful of channel capacity became almost a cliche. Various types of redundancies have been measured of different subject matters, and indicated the truthfulness of that statement.[1-3] A method proposed by Gouriet[1] provides a continuous indication of the entropy based on successive elements joint probability distribution, if an exponential distribution is assumed. The particular instrument called the "detail meter" differentiates the signal, provides an integrated value of the absolute value of the derivative, obtained by full wave rectification in a normalized form, where a maximum change at every picture element is called 100% detail. This instrument gives the maximum reading for a television signal corresponding to vertical bars (or a checkerboard) with the maximum resolution of the system (325 lines) but gives a reading of only 33% for a noise spectrum since the criteria of the exponential distribution is not satisfied. More significantly however when the instrument is connected to a video program line the reading is normally less than 3% and seldom is a reading in the order of 6% obtained. In the laboratory special subjects may be selected with substantially higher readings, however these scenes would not be considered pictures with entertainment value.

The detail meter thus indicated that there is a viewer preference of low detail or information content. Although some work in this field analyzed so called typical pictures we did consider to embark upon the project of finding either typical or preferable pictures. A clue was supplied by the psychologists, particularly by Prof. Y. LeGrand of Sorbonne, who performed some perception tests with images, which were to be sketched by an artist after a short exposure. These tests indicated that even after several second exposures only the contours are perceived. This is of course simply in line with the gestalt recognition of subject.

At this point the problem was then to establish a relationship between the gestalt concept and some unit familiar to the communication engineer; to provide a display which exposes the subject for a variable but precisely measured time and to find a suitable way for obtaining a measure of the observation.

One may consider each letter of the alphabet, or each number a gestalt and then knowing that random letters correspond to 4.7 bits per symbol, a relationship between a class of subject and the gestalt may be established. The perception of random letters however cannot be measured too easily and they form a rather limited class of gestalts. Nevertheless some reading speed tests and short exposure tests were performed and some of these results will be reported upon at another time.

The Test Setup

In order to control the exposure period precisely, an electronic switching

system was developed to switch video signals. The switching gear operates from the video synchronizing pulses, switching from one signal to another at the beginning of a field and switching back to the first signal again after 2,4,8,16,32,64 or 128 fields (or correspondingly after approx. 1/30, 1/15, 2/15, 1/4, 1/2, 1 and 2 seconds). The exposure time to the second signal can be set beforehand and the changeovers can be actuated when the observers are ready by a switch.

The use of a steady image, which is interrupted by the test image was adopted with the idea that the reappearing steady image does not permit the storage of the test image on the retina. This effect could be verified easily by using a uniform, white gray or black field as the steady image. In most of the tests the picture shown in Fig. 1 was used as the steady image.

The display device used was a TV station monitor with a 12 inch kinescope. The observers were asked to view the screen as they would view a television receiver. Generally a distance from 5 to 7 of the picture height was used. As a preliminary experiment the slide shown in Fig. 2 was flashed on for about twenty observers and they were asked to recite what they have seen as soon after being exposed to the picture as possible and a stenographer took down their remarks. For two fields of exposure (1/30 second), the observers noticed a change but could not tell what they saw. For an exposure of four fields they recognized either the horse or the house but not both. In 16 fields (approx. 1/4 sec.), they observed two connected objects such as: "a white horse pulling something", or "house and trees", "house with windows", and even reported inventions such as "a pony with a rider", "a house and a street and I think there was a car in the street".

In connection with these experiments it was already notable that (1) the center of the picture was perceived first; the majority of the observers, after recognizing one object, tried to name more by conjecture; (3) the number of objects recognized by the different observers for the same periods were remarkably uniform, and (4) a given total exposure seemed to yield the same amount of recognition whether it was taken in one exposure or in two exposures, each lasting half of the total period.

In the next experiment an attempt was made to evaluate the perception in bits and to eliminate any relationship between the recognizable objects. The slide shown in Fig. 3 was prepared and the evaluation of each object in bits was based on the following reasoning. Ogden[4] lists 200 picturable nouns among his 1000 basic words. Four of the five objects were among these, the pyramid however was not. Since it would have been a formidable job to count all the picturable nouns in a complete dictionary, a sampling technique was used and a factor of 4.7 was found for the letter B and a factor of 5.1 for the letter S, between the basic English list and the

Fig. 1

Fig. 2

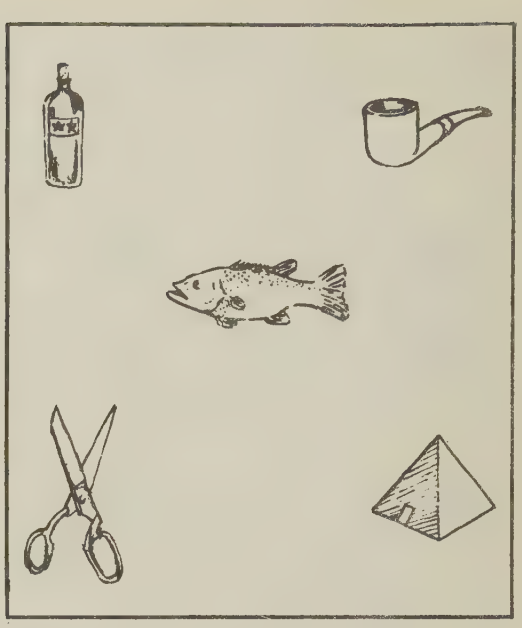


Fig. 3

Webster dictionary[6]. This provides an estimate of 1000 picturable nouns or for each gestalt approx. 10 bits of information ($\log_2 1024$). On the basis of this estimate then it was possible to test the perception speed in a bits per second scale.

The procedure of the test with slide 3 was the same as with the previous slide. The first tests with 8 fields (approx. 1/8 second) was inadequate for all observers to recognize any of the objects. A typical answer was, "approx. eight small figures in three columns that appeared to be line drawings of mythical animals". With an exposure of 16 fields (approx. 1/4 sec.) all observers recognized the fish. The test with an exposure period of 32 fields (approx. 1/2 sec.) was performed with 50 observers including engineers, stenographers, janitors, etc. Only one out of the 50 could name three objects. With one exception the rest named the fish and one more object, the one exception named the pipe and the scissors in that order. In all the tests with any exposure periods (about 120 tests altogether), if an object was recognized at all, the center figure was the first named.

The recognition of two independent objects corresponds to 20 bits and since the exposure was 1/2 second, it corresponds to 40 bits per second perception speed. It may be interesting to compare this figure with the capacity of a television signal channel. The number of different messages that can be constructed with 200,000 picture elements per frame with 100 distinguishable brightness levels is $10^{400,000}$. On the basis of the relationship of

$$2^X = 10^{400,000}$$

$$X = \frac{400,000}{\log_{10} 2}$$

this corresponds to approx. 1.3 million bits per frame or 40 million bits per second.

It was felt that once the technique described above was developed, the perception speed should be verified with gestalts of smaller classes, and therefore shorter periods. One experiment with a slide which had objects with names all starting with the letter B, failed to produce a greater object recognition higher than two per 32 fields after the observers were instructed about the restriction. On discussing the observation with some of the viewers, it was evident that this type of restriction imposes a difficult mental process, which apparently causes a slowing down in the perception speed.

Another experiment in which the observer was shown the chart of 16 symbols shown in Fig. 4 beforehand and then asked to recognize one of these symbols, yielded excellent consistency between tests and the results obtained with the general class of picturable nouns. By flashing the single symbol (corresponding to four bits) for 4 fields (1/15 of a second) therefore 60 bits per second only one out of twenty observers guessed the correct symbol, which might even be accounted for on the basis of chance. When the exposure period was doubled (8 fields,

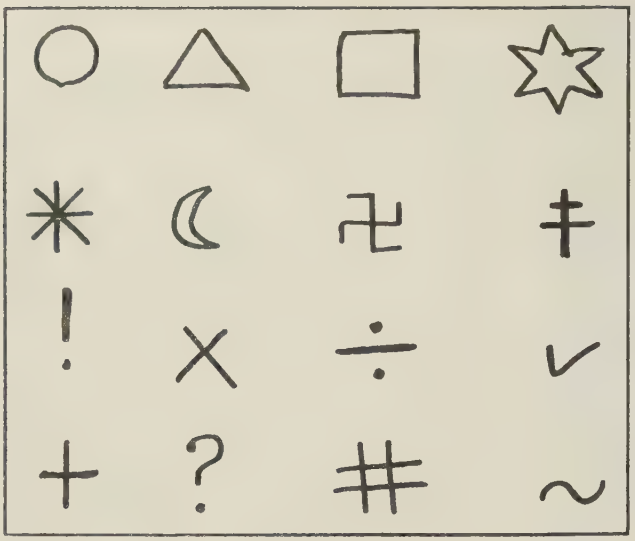


Fig. 4

2/15 sec.) and therefore the perception rate demand was reduced to 30 bits per second, only one of another batch of 20 observers failed to recognize the right symbol.

The importance of the steady picture before and after the exposure of the unknown picture was demonstrated easily by the use of the symbol test. While with the steady picture, the symbol could not be recognized in 8 fields; if the exposure to the symbol was followed by a blank screen, even one field (1/60 of a second) was enough to pick the right symbol. Based on these tests it is believed, that the type of tests, where an image is flashed on and off by optical means, the retinal retention is tested rather than the perception speed.

One further test involved the instruction of the observer that the symbol to be shown is a choice of two of the symbols from Fig. 4, thus each recognition corresponding to 1 bit of information received correctly. The speed for this test was set at 1/30 of a second and we expected to perform a large number of tests in order to get an average, however the 30 bits per second provided perfect recognition every time and accordingly ten tests were considered conclusive.

The author gratefully acknowledges the help of Mr. R. Staffin with the tests. Mr. Staffin also constructed the timing switch.

References

[1] Gouriet, G.G., "A Method of Measuring Television Picture Detail", Electronic Engineering, Vol. 24, 1952, p. 308.

[2] E.R. Kretzmer, "Statistics of Television Signals", BSTJ, Vol. 31, 1952, p. 751.

[3] C.W. Harrison, "Experiments with Linear Prediction in Television", BSTJ, Vol. 31, 1952, p. 764.

[4] C.K. Ogden: The Basic Words, Kegan Paul, Trench and Co., London, 1933.

[5] Webster's New Collegiate Dictionary, G and C. Merriam Co., 1953.

HUMAN MEMORY AND THE STORAGE OF INFORMATION*

George A. Miller

Harvard University
Cambridge, Massachusetts

Abstract

The amount of selective information in a message can be increased either by increasing the variety of the symbols from which it is composed or by increasing the length of the message. Psychological experiments indicate that the variety of the symbols is far less important than the length of the message in controlling what human subjects are able to remember. Two messages equal in length but differing in the amount of information per symbol are equally easy to memorize. This fact provides an opportunity for the effective use of recoding procedures and reveals the mental economy involved in organizing the materials we want to remember.

An apparent exception to the rule that length, not variety, is the limiting factor in human memory occurs in the case of redundant messages. If two messages of the same length differ because one contains redundancy familiar to the learner and the other does not, the redundant message will usually be easier to learn and remember. In terms of the theory of information, redundancy can be viewed equally well as a reduction in the information per symbol or as a reduction in the effective length of the message. Psychologically, however, these two alternatives are not equivalent; redundancy permits a reorganization into familiar sequences in a way that effectively shortens the length of the message and so makes it easier to memorize, but this is not psychologically equivalent to reducing the amount of information per symbol.

It is as if each storage register could accept any one of a tremendous variety of alternative symbols, but the number of registers available was quite limited. If we use these registers to store binary symbols, the storage is inefficient. If we group the binary symbols into sequences, give each sequence a different name, and store the recoded names, we can make much more efficient use of the registers. Familiar redundancy is helpful because it enables us to recode more efficiently.

These results for human memory are all the more striking in view of the fact that the amount of information per symbol is a critically important variable controlling the accuracy of our perceptions.

* Preparation of this article was supported in part under Contracts AF 33(038)-14343, AFCRC TR 56-54 and Nonr 1866(15) (Project NR142-201, Report PNR-185). Reproduction for any purpose of the U. S. Government is permitted.

Introduction

The development of a mathematical theory of communication has stimulated considerable interest among experimental psychologists in the measurement of human capacities to process selective information (8). If a human operator is regarded as a communication channel with stimuli for inputs and responses for outputs, it is possible to estimate maximum rates of transmission through him. The amount of input information can be varied either by increasing the rate of presentation of the stimuli or by increasing the amount of information per stimulus. Or, if a human operator is looked upon as a storage device, one can determine the conditions under which he learns or remembers the largest amounts of selective information. Most of these applications have used the discrete model developed by Shannon (13), since psychologists tend to regard stimuli and responses as discrete events; however, the continuous case can be used in the analysis of tracking behavior.

Rates of Transmission

Studies of the maximum rate of transmission yield values so low that one is forced to conclude that the human nervous system was not designed with transmission as its major objective. The practical upper limit under optimal

conditions seems to be around 25 bits/second, and 40 bits/second is as high as anyone has yet claimed to be able to achieve. According to Quastler (12, pp. 341-349), there are two factors that limit the maximum rate. The first arises from the fact that people cannot respond more than about nine times/second, and this is effectively reduced to five responses/second when discrimination is required. A second limit intrudes because people cannot discriminate rapidly among more than about 32 equi-probable alternatives without becoming confused. If the task exceeds either of these limits, performance deteriorates rapidly. In these studies Quastler used skilled typists and pianists and concluded that 25 bits/second was the practical upper limit for typing and playing the piano.

Fitts (4), using less familiar tasks like tapping alternately two targets, transferring discs from one pin to another, and transferring pins from one set of holes to another, also found a consistent upper limit to the amount of information that people could generate, but the limit was between 10 and 12 bits/second. The difference between Quastler's limit of 25 bits/second and Fitts' limit of 12 bits/second is due to the fact that Quastler's tasks used both arm and finger movements, whereas Fitts' tasks involved only arm movements. Still different limits would probably result for other muscle groups. Such differences indicate that we must be careful to specify what segment of the total organism is involved in producing the responses before our statements of channel capacity are meaningful.

Since the muscles are physically capable of responding much faster than ten contractions/second, the limits observed in these experiments are presumably attributable to the fact that excitation and inhibition cannot be made to alternate rapidly in the central nervous system without losing precise control of the movement. The observed trading relation between speed and accuracy which results in the constancy of the informational measure is the result of the time taken for central organizing processes to occur.

The inference that the bottleneck occurs centrally is supported by the fact that the rate of transmission of selective information is greatly reduced if there is an unnatural or unfamiliar relation between the stimulus and the response. For example, Fitts and Deininger (5) found that if the response was a movement toward a target light at the instant the light appeared, performance was far superior to the case in which an arbitrary direction of movement was assigned for each target light. This superiority persists even after extended practice (6). Presumably the arbitrary relation between stimulus and response requires an additional recoding operation by the organism and the more complex this central transformation becomes, the slower the rate of transmission. Thus, the limit depends not only upon the discriminability of the stimuli and upon the particular response system employed, but also upon the degree of congruence between stimuli and responses. Fitts has referred to this fact as the principle of "stimulus-response compatibility."

Span of Absolute Judgment

The limitations imposed on our capacity to discriminate among different stimuli can be studied in more detail if we remove the requirement that the operator must respond as quickly as possible. That is to say, we give the person unlimited time in which to respond, but ask him to be as accurate as possible. A stimulus is presented and the observer is required to identify which one of a set of alternative stimuli it is. Under these conditions of judgment it appears that the complexity of the stimulus is an extremely important factor. If we take the simplest possible stimuli -- pure tones, monochromatic lights -- and vary them along a single dimension, subjects can identify accurately from 5 to 15 alternatives. For such unidimensional stimuli, therefore, the discriminative capacity ranges between two and four bits/judgment. However, if the stimuli differ from one another in several dimensions, discriminations can be made simultaneously along each dimension. The total discriminative capacity is increased by increasing the number of dimensions of variation in the stimuli, but the accuracy of judgment on each individual dimension decreases -- there is a slight interference or "masking" effect in multidimensional discriminations. With sounds that differed from one another on six different acoustic dimensions, for example, Pollack and Ficks (11) obtained the value of 7.2 bits/judgment. This is a large increase over the three bits/judgment obtained with unidimensional stimuli, but it represents a relative decrease to 1.2 bits/judgment/dimension.

Apparently we are designed to operate best when we must make relatively crude judgments of several attributes simultaneously. We are not able to make extremely precise identifications along a single dimension. Miller (9) has referred to this limitation as defining a "span of absolute judgment."

This perceptual limitation is surprising when we remember that people can detect very small differences between two stimuli presented at the same time or in immediate succession. Relative judgments are far easier than absolute identifications. The limit does not seem to reside peripherally in the receptor organs, but in a central process of judgment that is involved when we attempt to identify or name a particular, individual stimulus. We are far more accurate in saying that two stimuli are different than we are in saying which particular stimuli they happen to be. Thus we are once again, as in the case of the muscles, forced to look to the central nervous system for the cause of our limited capacity to process information.

Span of Immediate Memory

Once we have removed the instruction that the person must respond as rapidly as possible, it is quite natural to go one step further and ask the person to withhold his response until several stimuli have been presented in succession. Thus we move from measuring bits/second to bits/judgment and on to bits/sequence. To discrimination and response, we add the task of storage. The limitations of memory, however, appear to be a different sort than the limitations of discrimination or response.

In the simplest test of mnemonic capacity, a sequence of symbols (usually decimal digits) is read aloud or shown to the person at a regular rate (usually one per second) and at the end of the sequence he is asked to repeat or write the symbols in the correct order. The experimenter begins with short sequences and increases the length until the person is no longer able to repeat the entire sequence without error. This point is called the "span of immediate memory." The procedure was adopted by Binet in his first scale for measuring mental age and has been retained in all subsequent revisions of the scale. The memory span is not a perfect measure of intelligence, however, since a long span does not necessarily indicate high intelligence. It has been retained because an unusually short span is a reliable indicator of mental deficiency.

Inasmuch as the amount of information required to specify the stimulus or to control the response sets the perceptual and motor limits, it is natural to ask whether the span of immediate memory shows a similar invariance when information measures are applied. The answer is unequivocally negative. It is the length of the sequence, rather than the amount of information per item, that is the critical factor. For example, a person who can repeat nine binary digits will have a span of about eight decimal digits, seven letters of the alphabet, or five monosyllabic English words. These represent 9, 25, 33, and about 50 bits, respectively. Clearly, the memory span is more nearly invariant when we measure it in terms of the length of the sequence than when we measure it in terms of the amount of information stored. Whereas perception and response are limited by the number of bits of information, immediate memory is limited by the number of items or "chunks" of information (9).

This fact leads to an insight into the economics of cognitive organization. Since it is as easy to remember a lot of information (when the items are informationally rich) as it is to remember a little information (when the items are informationally impoverished), it is economical to organize the material into rich chunks. To draw a rather farfetched analogy, it is as if we had to carry all our money in a purse that could contain only seven coins. It doesn't matter to the purse, however, whether these coins are pennies or silver dollars. The process of organizing and reorganizing is a pervasive human trait, and it is motivated, at least in part, by an attempt to make the best possible use of our mnemonic capacity.

Recoding

The effect of reorganizing, or recoding, the input can be illustrated by a trick that computer engineers use to remember a long sequence of binary digits. The sequence of binary digits is first grouped into successive triplets and then each one of the eight possible triplets is translated into a single octal digit. For example the sequence 010111001001000110 is grouped 010-111-001-001-000-110 and recoded into 271106. The original 18 binary digits far exceed the span of immediate memory, but the six recoded octal digits are easily remembered. After a little study of the binary-to-octal transformation, the engineers are able to deal with almost three times as much information as before.

A great deal of our learning concerns the development of these rules for reorganizing the input information. For instance, when a man first begins to receive Morse Code, each dit and dah seems to be an isolated item of information and he gets lost if he falls more than two or three letters (five to ten dits and dahs) behind the transmitter. As he learns to recode the dits and dahs into letters, then words, and

later phrases, he is able to deal with larger and larger chunks of the message. An experienced operator may sometimes fall as much as ten words behind a familiar message without becoming lost; ten words represent about 150 dits and dahs. The experienced operator is able to store that sequence of 150 items away in memory because he has organized them into familiar groups in much the same way computer engineers organize binary into octal digits.

It is possible to think of many of the great advances in human thought as discoveries of more economical ways to package information that must be stored in the mnemonic warehouse. The superiority of the Arabic over the Roman notation for numbers is a case in point, and many advances in mathematics--the calculus, matrix algebra, statistical theory, and probably many others--can be considered as providing a mnemonically better set of symbols for representing the relevant aspects of a problem in such a way that we can grasp them in a single act of thought. It would be foolish to argue that mental economy was the only result of mathematical inventions, but certainly the discovery of a good system of notation is an important step toward new insight. It is interesting to speculate whether the construction of large computing machines with enough storage to make such recoding unnecessary will have any important effect upon mathematical creativity. Our growing capacity to solve problems by electronic computation before we have properly understood them has disturbed many pure mathematicians.

It is also possible to argue that the value of natural laws is at least in part due to the fact that they summarize in a convenient formula a tremendous amount of information collected from individual observations in numerous experiments. The law of gravitation or the gas laws, for example, might just as well be summarized in tables with the experimental measurements tabulated under the appropriate experimental conditions; such tables would have the tremendous disadvantage that, although they contain exactly the same information, we could not apprehend simultaneously all that they have to tell us.

A less precise but more general technique for organizing our experience into convenient units is provided by language (7, pp. 223-237). When you witness a scene or hear a story that you want to remember, you try to translate it "into your own words," into the linguistic units that will fit into your own cognitive hierarchy. This highly schematic, verbalized abbreviation is remembered. Then when you try to recall you must decode. Since the fit of words to experience is seldom as tight as the fit of laws to data, the decoding process often goes astray. You supply details by secondary elaboration that are consistent with your coded memory. Often these details are wrong. Psychologists have been interested in such systematic distortions, because they become of practical consequence for legal testimony and the propagation of rumor (1).

Rote Memorization

These speculations indicate some of the implications of the fact that it is length, not variety, that imposes the major restriction upon immediate memory. Immediate memory, however, is merely the simplest of our mnemonic functions. What happens when the amount of material that we must deal with greatly exceeds the span of immediate memory?

One way that psychologists have studied how people assimilate large quantities of material is to ask them to memorize it and to record how much of the material they have mastered after each rehearsal. This procedure has the merit that it provides a reasonably objective quantification of the progress of learning. It can be used either with meaningful, connected text or with nonsensical, disconnected sequences. We shall consider both cases, taking first the simpler, though less natural, case in which the material to be learned consists of a sequence of symbols chosen independently from a given set of alternatives.

Dr. Sidney L. Smith and I asked subjects to memorize long sequences of items in which each item was chosen from among either 2, 8, or 32 alternatives (0 or 1, 0 through 7, or all the letters of the alphabet except Q plus the numerals 3 through 9). Lists containing 10, 20, 30, and 50 successive items were constructed for all three kinds of test material. Each list was presented to the learner a symbol at a time at a rate of one symbol/second and after the complete presentation he wrote down in the correct order as much of the list as he could remember. The same list was presented repeatedly until he was able to reproduce it all without error. Six people learned the entire set of lists.

The lists were constructed in such a way that each item contained 1, 3, or 5 bits of information. If the difficulty of the task depends upon the amount of information involved, it should take just as long to memorize a sequence of ten items selected from 32 alternatives as it takes to memorize a list of 50 items selected

from two alternatives. The averages of the number of trials required in these two cases, however, were 2.5 and 12.2, respectively, a very reliable difference. Thus we can safely say that, even when the material to be remembered exceeds the span of immediate memory, the difficulty of the task does not depend upon the amount of information that the material contains.

If the difficulty of the task depends upon the number of items involved, it should take just as long to memorize a sequence of N items selected from 32 alternatives as from eight or two alternatives. This prediction is much closer to the facts. The number of repetitions required to memorize the 32-alternative and 8-alternative lists were not significantly different for any of the lengths of list used. The sequence of binary items, however, was slightly easier and required about 20% fewer repetitions; this discrepancy is caused presumably by the fact that binary sequences are especially easy to group and recode. In general, however, even when the material to be remembered exceeds the span of immediate memory, the difficulty of the task depends critically upon the length of the list.

To summarize, therefore, we may say that when the material to be learned does not form a familiar sequence, the difficulty depends primarily upon the length of the material and is relatively independent of the amount of information it contains. Under these conditions, it is just as easy to memorize a lot of information as to memorize little information. And this conclusion holds both for the span of immediate memory and for materials that exceed the span and must be repeated several times. Our confidence in this generalization is supported by the fact that Brogden and Schmidt (2, 3), working with quite different techniques and unaware of the hypothesis Smith and I were trying to test, obtained very similar results.

The Unitization Hypothesis

The fact that there is a limited span of immediate memory poses a paradox for psychologists; one might almost call it the central issue for studies of verbal learning. If we can remember seven items without any trouble, why can't we simply hold those seven and take on seven more? What is it about the second seven that makes us forget the first seven? A variety of explanations have been proposed. "Leaky bucket" hypotheses hold that a point is reached at which the old material leaks out as fast as the new is put in; that the memory traces fade away until we start to forget as fast as we learn. "Cross talk" hypotheses hold that a point is reached at which the separate signals begin to interfere with one another; that we have competing tendencies to respond with different items instead of the correct item. "Sabotage" hypotheses hold that a point is reached at which new items begin to wreck the established mnemonic machinery; that each item sets up an active inhibitory process that accumulates until it exceeds a critical level. "Standing room only" hypotheses hold that there are only so many seats available in the mental amphitheater; that there is not time to organize the material properly into supraordinate units in order to fit it into the available number of slots.

Each of these several varieties of hypotheses has implications for a wide range of phenomena that have been observed in studies of verbal learning. This is not the place to develop them or to compare their relative merits. They are mentioned here only to indicate that there is some divergence of opinion among psychologists and that the opinions of this author are not likely to represent the final word on this complex topic. The present exposition of the unitization, or "standing room only," hypothesis, therefore, should be read with only one eye.

Suppose we take quite literally the assumption that our memory is capable of dealing with only seven items at a time. It is as if we were dealing with a computing machine that has a small, fixed number of storage registers. Each register can accept any of a tremendous variety of different symbols, so the total amount of information that can be stored is quite large. The design of this storage system, however, makes it necessary to recode the input in order to reduce it to a small enough number of symbols. When the number of items in the input exceeds the number of registers, therefore, the items must be grouped according to some scheme of organization, new symbols must be chosen to represent the new groups, and these recoded symbols are then stored in the registers. The learning process consists largely of reorganizing and symbolizing the input until it is reorganized into supraordinate units sufficiently simple to fit into the machine. It is this grouping and naming that we shall call "unitization."

When a person submits himself to a psychologist who asks him to memorize some stupid and useless sequence of symbols, he probably unitizes the material in an *ad hoc* manner that is quite tentative and transient, but is adequate for the immediate purposes. When he sets out to learn something that he is personally inter-

ested in and that he expects to have use for, however, he is probably much more careful to organize the material in a way that fits well into his established cognitive structure. Without the pressure of time, he can explore various alternative unitizations until he finds one that works best for him and promises the best recall at any later date. In either case, however, his task is to create a hierarchy of units in such a way that by recalling the few, informationally rich and suggestive units at the top of the hierarchy he can then recover the more numerous, more detailed items at the bottom.

The importance of this organizing process is introspectively obvious, but it is quite difficult to get at experimentally. The behavioral system that shows the clearest hierarchical organization of small units into larger, supraordinate units is, of course, language. Ever since Plato observed that thought is the soul's discourse with itself we have been aware of the intimate relation between thinking and talking. Although we are far from understanding the precise nature of this relation, it does seem reasonable to assume that the hierarchical organization of language--sounds or letters, syllables, words, phrases, clauses, sentences, paragraphs--is not an accidental pattern, but truly represents the preferred mode of operation of our mental machinery. This is not to suggest that the laws of grammar are the laws of thought; the fallacy of this argument is exposed by the variety of grammatical laws in different languages. But the existence of some kind of hierarchical organization in all languages must not be ignored. The identification of the appropriate units would be as significant for psychology as the molecular theory was for physics and chemistry or the cellular theory for biology.

Before the "standing room only" hypothesis can be of much value to us, however, we must somehow derive from it quantitative predictions that can be tested by experiments which psychologists are able to conduct. By way of illustration, we might develop some of the following as theorems from a basic postulate: (1) Anything that interferes with the recoding process will interfere with memory. (2) Since the rate of presentation of the material can be too rapid to permit organization to occur, we would predict that a slow rate of presentation would lead to learning in fewer repetitions than would a fast rate. Further (3) the superiority of the slower rate would tend to disappear as the material became more meaningful, since the organization into units is easier with familiar text than with random sequences of symbols. Also (4) since connected discourse fits into a scheme of organization that we have already learned, it should be much easier to learn than the same amount of nonsense. (5) The amount of organizing required increases as the length of the material increases, other things being equal, and thus long passages are harder to learn than short ones. (6) The process of organizing should begin at some clear focus or reference point in the material, usually the beginning or the end, so that the middle of the sequence should be learned last. (7) The particular confusions and mistakes that occur should be predictable from a knowledge of the method of recoding that was used. (8) The learner's expectations will influence the way he organizes, so that tests of retention which violate his expectations should lead to poor performance. All of these predictions are supported by experimental evidence, but that fact is not decisive, because they can also be accounted for in terms of other theories.

Meaningful Materials

A major difficulty encountered in any attempt to apply the unitization hypothesis to particular experimental data is the specification of the supraordinate units that the learners are using. It is fine to know that it is length rather than variety, chunks rather than bits, that limits our memories. But this conclusion would be more useful if we had a better way to recognize what size chunks were used. For example, we can usually repeat a 20-word sentence after hearing it once. How many items--100 letters, 30 syllables, 20 words, 6 phrases, 2 clauses, or one sentence--does this sentence contain? We know that it contains about 120 bits of information, because we have a definition of the bit that is independent of our subjective organization of the sentence. But the essence of the chunk is that it is imposed by the person. For example, someone who knew nothing of English except the alphabet would have to treat the sentence as if it were 100 units long, whereas someone who knows English well might deal with it as if it were six units long. We cannot define the unit of organization independently of the learner.

One way to get at the problem of defining the unit is to use sequences of words constructed at different orders of approximation to English. The 0-order approximation is constructed by selecting words at random from a dictionary, so that each word has an equal chance to occur in the sequence. The 1-order approximation consists of words chosen randomly from connected text, so that each word occurs according

to its natural probability in English discourse. The 2-order approximation consists of words selected in the context of one preceding word, so that each successive (overlapping) pair of words occurs with its natural probability in English discourse. One way to construct such sequences is to select a word at random and search through a text until this word occurs. Then take the word which follows it and search until this word occurs again. Then take the word that follows it, etc., continuing in this way until a passage of the desired length is obtained. An alternative method is to ask a person to use the word in a sentence. Then take the word that follows it in his sentence and give this word to another person to use in a sentence. Then take the following word in his sentence, etc., until the passage is completed. The 3-order approximation consists of words selected in the context of two preceding words and, in general, the N-order approximation contains words chosen in the context of N - 1 preceding words. As N increases the sequences begin to sound more and more like English. The following are some examples:

0: Betwixt trumpeter pebbly complication vigorous tipple careen obscure attractive consequences expedition pane unpunished prominence chest sweetly basin awoke photographer ungrateful

1: Is to went biped the of before love turtledoves the spins and I of yard than ask went Greek yesterday

2: Sun was nice dormitory is I like chocolate cake but I think that book is he wants to school there

3: Family was large dark animal came roaring down the middle of my friends love books passionately every kiss is fine

5: Road in the country was insane especially in dreary rooms where they have some books to to buy for studying Greek

7: Said that he was afraid of dogs marked with white spots and with black spots covering it the leopard did

Miller and Selfridge (10) constructed 0, 1, 2, 3, 4, 5, and 7-order approximations to study how well such materials could be recalled. At each order of approximation, lists 10, 20, 30, and 50 words in length were used. A group of people heard the words read aloud once in a regular monotone and attempted to recall them immediately afterward. As might be expected, it was found that the higher-order approximations were remembered best. The interesting result, however, was that most of the improvement had occurred by the time a 5-order approximation was reached. The introduction of contextual constraints extending beyond about five words added little to facilitate recall. It is the short-term dependencies that are most important. A 5-order approximation is still nonsense. Apparently it is not important that the sequence of words have a meaning. It is sufficient that it does not violate familiar intraverbal connections extending over sequences of only four or five words. On the basis of this evidence, therefore, we might argue that the natural size unit for dealing with connected text averages about five words in length.

Miller and Selfridge also included randomly selected segments of connected texts in their tests and found only a slight, insignificant difference in recall for 5-order, 7-order, and textual materials. Apparently this result was due to an unfortunate selection of textual materials, for other experimenters have found connected text slightly easier to learn. It seems that meaning does add something to recall that is not contained in the nonsensical approximations. This fact was clearly demonstrated by Marvin Levine in an unpublished study. Levine

used short anecdotes written to have exactly the same length as the other passages. The meaningful unity of the anecdote produced recall scores about 10% higher than he obtained with 7-order approximations. Thus larger units can play a significant role in recall. Nevertheless, the fact remains that it is the short-term connections that produce the major effects.

Redundancy

If we view the results of the Miller-Selfridge experiment in the light of the preceding discussion, they seem at first glance to violate our general thesis that it is length, not variety, that makes a sequence of symbols difficult to memorize. The 0-order approximation is less redundant and so contains more information per word than does the 7-order approximation, and it is correspondingly more difficult to remember. Although the data do not permit a precise calculation, it is approximately true that the same amount of information was recalled with 50-word passages independent of the order of approximation to English. If these were the only data available, therefore, we might conclude that it is the amount of information, not the length, that is the critical variable.

However, the Miller-Selfridge results do not give a decisive answer to this question. It is indeed true that the 0-order approximation contains more information per word, but it is also true that it is psychologically longer than the same number of words in a 7-order approximation.

That is to say, each word in a 0-order approximation is a separate unit, whereas successive words in a 7-order approximation can be grouped into familiar, supraordinate units. Thus the experiment confounds the two variables, amount of information and number of units, that we are attempting to separate. On the basis of these data alone we cannot say conclusively whether it is the greater number of units or the greater number of bits that makes the 0-order approximation hard to recall.

Shannon (14) has estimated that English is about 75% redundant. This redundancy provides a margin of safety for our perceptual and motor processes. It enables us to mis-speak or to mis-hear the details of a message and to correct our errors on the basis of context. For the purposes of remembering, however, this redundancy would seem to be very inefficient. If it is length that burdens our storage facilities, why do we deliberately make our messages four times as long as would be necessary if we used our alphabet with maximum efficiency?

This question springs from an easy confusion between information theory and psychological theory. Within the framework of information theory it is true that redundancy can be treated either as (1) a reduction in the information per symbol, or (2) a reduction in the effective length of the message. Psychologically, however, these two alternatives are not equivalent; the fact that the message is relatively predictable enables a man to organize it into supraordinate units, and it is these organized units, corresponding roughly to what we call "ideas," that he stores away in memory. Information theory does not say that 75% of our ideas are redundant (though this may indeed be the case), but only that the same ideas could be encoded in 25% as many letters. In terms of psychological units, therefore, our messages are not four times as long as necessary to the person familiar with the organization of the language.

On the basis of the available evidence it seems that memory, unlike sensory-motor processes, is not limited primarily by the amount of information it must process. Memory is limited by the number of psychological units it must handle; under appropriate recoding transformations these units can come to represent large amounts of information. In one sense, this is a disappointing conclusion. Research would be easier if the well-defined bit could be used to measure the capacity of the memory. Unfortunately for psychologists, the human organism was not designed for the convenience of researchers. The problem of defining the size of the chunk of information that people treat as a unit has not been solved by information theory, and the psychologist's task remains as difficult as ever. But on the comforting side of the picture, information theory has helped to clarify the purposes behind our persistent classifying and unitizing of experience. It is informationally profitable, both for communication and memory, to organize and symbolize. For this reason, the measurement of selective information has become an important tool in psychological research.

References

1. Allport, G. W., and L. J. Postman. The Psychology of Rumor. New York: Holt, 1947.

2. Brogden, W. J., and R. E. Schmidt. The effect of number of choices per unit of a verbal maze on learning and serial position errors. J. exp. Psychol., 1954, 47, 235-240.

3. Brogden, W. J., and R. E. Schmidt. Acquisition of a 24-unit verbal maze as a function of the number of alternative choices per unit. J. exp. Psychol., 1954, 48, 335-338.

4. Fitts, P. M. The information capacity of the human motor system in controlling the amplitude of movement. J. exp. Psychol., 1954, 47, 381-391.

5. Fitts, P. M., and R. L. Deininger. S-R compatibility: Correspondence among paired elements within stimulus and response codes. J. exp. Psychol., 1954, 48, 483-492.

6. Fitts, P. M., and C. M. Seeger. S-R compatibility: Spatial characteristics of stimulus and response codes. J. exp. Psychol., 1953, 46, 199-210.

7. Miller, G. A. Language and Communication. New York: McGraw-Hill, 1951.

8. Miller, G. A. What is information measurement? Amer. Psychologist, 1953, 8, 3-11.

9. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev., 1956, 63, 81-97.

10. Miller, G. A., and J. A. Selfridge. Verbal context and the recall of meaningful material. Amer. J. Psychol., 1950, 63, 176-185.

11. Pollack, I., and L. Ficks. Information of elementary multi-dimensional auditory displays. J. acoust. Soc. Amer., 1954, 26, 155-158.

12. Quastler, H. (Ed.) Information Theory in Psychology. Glencoe: Free Press, 1955.

13. Shannon, C. E. A mathematical theory of communication. Bell Syst. Tech. J., 1948, 27, 379-423.

14. Shannon, C. E. Prediction and entropy of printed English. Bell Syst. Tech. J., 1951, 30, 50-64.

THE HUMAN USE OF INFORMATION
III. DECISION-MAKING IN SIGNAL DETECTION AND RECOGNITION SITUATIONS INVOLVING MULTIPLE ALTERNATIVES*

John A. Swets and Theodore G. Birdsall
University of Michigan

Summary

A general theory of signal detectability, constructed after the model provided by decision theory, is applied to the performance of the human observer faced with the problem of choosing among multiple signal alternatives on the basis of a fixed, finite observation interval. An extension of the theory, previously reported, of deciding among two alternatives, is developed in detail, permitting the treatment of the two simple cases involving multiple alternatives that are studied experimentally.

In both cases, a priori probabilities are assigned to the occurrence of the relevant signal alternatives, and values are assigned to the possible decision outcomes. This assignment permits definition of the expected value of a decision, and specifies as the optimum decision criteria those that maximize expected value. The experiments are primarily concerned with determining the ability of the human observer to successively establish optimum decision criteria in accordance with changes in the a priori probabilities and risk functions.

The experimental results are portrayed in the form of comparisons of the payoff obtained by the observer and the theoretically maximum payoff attainable, and comparisons of the response-frequency tables generated by the theory and by the observer. The results indicate that a highly simplified theory is adequate for prediction of the obtained payoff and response-frequency tables to within a few percent. They also indicate the fairly large extent to which intelligence may influence a sensory process usually assumed to involve fixed parameters.

Introduction

Several experiments previously reported have supported the applicability of the model provided by the theory of statistical decision, or the theory of testing statistical hypotheses, to the behavior of the human observer in signal detection and recognition situations. In particular, a theory which assumes the human observer to be capable of adapting to any of a number of optimum decision rules has been shown experimentally to constitute an adequate description of his behavior in signal reception problems involving two signal alternatives or hypotheses.

* This work was sponsored by the U. S. Army Signal Corps.

This paper is concerned with more complex situations. It reports an extension of the decision-making theory of detection in which optimum behavior is specified for two situations, each involving three signal hypotheses. The results of experimental tests of the extended theory, employing auditory signals embedded in noise, are also presented.

The relevance of the theory of statistical decision to the general theory of signal detectability has been discussed in other papers [3,14,15]. The application of the mathematical theory of signal dectability to the behavior of the human observer in several different detection and recognition problems, involving both visual and auditory signals, has also been the subject of a series of papers [1,6,8,9,10,11,12,13]. The next section summarizes those aspects of the papers referred to that are basic to the experiments reported here. In addition, the next section elaborates for the first time the theory specific to the present experiments. A more general treatment of the multiple-hypothesis theory is that of Middleton and Van Meter [15]. The development of the next section is somewhat simpler in that it deals only with the specific three-hypothesis cases studied experimentally.

The Decision-Making Theory of Detection

Two assumptions are primary in the application of the theory of signal detectability to the behavior of the human observer. The first is that sensory systems are basically communication channels, conveying information about environmental events to higher centers. It is supposed that, at the higher centers, this sensory information combines with a priori information to serve as the basis for decisions about the events initiating the sensory information. The second assumption is that the sensory systems are noisy channels, that is, that they themselves continuously generate irrelevant, random activity that is not easily distinguishable from the activity initiated externally. In other words, partially degraded information about immediate environmental events is displayed at the higher centers; here the display is observed and a decision is made.

The Theory of Testing Two Hypotheses

In the fundamental detection problem, a function of a fixed time interval is observed. The observer, in effect, is asked to decide whether the observation arose from noise alone or from signal-plus-noise, where the signal is known to be from a certain ensemble of signals. In terms of actual practice, the observer is asked to

state, after each presentation of the fixed time interval, which may or may not have contained a signal, either "yes, a signal was present" or "no, no signal was present."

It is assumed that the space of all possible observations contains families of nested sets, which are here called "criteria." For a given signal ensemble and type of noise only one of these families is employed. In a specific situation, the "yes" or "no" response after a particular observation is determined entirely by whether or not the corrupted observation falls inside a specific criterion. A less frequent "yes" response is obtained by using a criterion which is a subset of the first one. Since the criteria of a given family are nested, they can be simply ordered, and therefore this ordering is viewed as contained in a linear continuum. Of course, a change of signal ensemble or noise type may change the family of criteria, and thus change the linear continuum. However, the assumption that the observations are in nested criteria is sufficient to allow the observations to be represented along a continuous axis.

As implied in the discussion of the preceding paragraphs, the observations resulting from noise alone are regarded as randomly distributed, as are the observations arising from a signal of a given strength plus noise. The mean of the signal-plus-noise distribution is assumed to be a monotonic increasing function of signal strength. If each observation may arise from noise alone, it is possible to distort the linear continuum, maintaining only order, so that the distribution of observations (on the continuum) is Gaussian with zero mean and unit variance. This must be qualified to allow gaps in the distorted continuum if, in the original, there were points of positive probability.

This conception, when the signal-plus-noise (SN) distribution is simply a translation of the noise-alone (N) distribution, is depicted in Figure 1. The abscissa represents the observation, x; the ordinate represents the probability that a given observation will result from noise alone and from signal-plus-noise. Although unnecessary, it may be helpful to conceive of the observation variable as some measure of neural activity. The noise assumption (neural activity independent of external stimulation) is consistent with present knowledge of neurophysiology.

The Definition of Criterion and Likelihood Ratio. According to the conceptual scheme of Figure 1, the observation variable is continuous, and any given observation may arise either from noise or signal-plus-noise. For this problem, the observer must establish a level of confidence, or a boundary of the criterion, and base his decision on the relation of the observation to this boundary. In terms of the theory, the observer chooses a set of observations (the criterion A) such that an observation in this set will lead him to Accept the hypothesis SN, that is, to say that a signal was present. All other observations are in the Complement of the criterion, CA; these are regarded by the observer as representing noise alone. The criterion A, with reference to Figure 1, consists of the values of x greater than some critical value.

If the number of possible observations is countable, the conditional probability that signal-plus-noise will yield a given observation is denoted $P_{SN}(x)$; the probability of this observation, given noise alone, is denoted $P_N(x)$. The ratio of these two probabilities defines the likelihood ratio, $\ell(x) = P_{SN}(x)/P_N(x)$. Here the observation space is considered continuous, and, accordingly, the probability density functions $f_{SN}(x)$ and $f_N(x)$ are used; $f_{SN}(x)$ corresponds to the curve labelled SN in Figure 1 and $f_N(x)$ is the curve labelled N. In this case, $\ell(x) = f_{SN}(x)/f_N(x)$. A decision concerning signal existance is reasonably based on this quantity, the relative likelihood that the observation x arose from signal-plus-noise and noise alone. According to the theory, for every x, the observer can estimate $\ell(x)$.

A criterion is conventionally evaluated in terms of the integrals of the density functions over the criterion A, since the integral of $f_{SN}(x)$ over A is the conditional probability of detection, $P_{SN}(A)$, and the integral of $f_N(x)$ over A is the conditional probability of a false alarm, $P_N(A)$. A plot of $P_{SN}(A)$ vs. $P_N(A)$ as the decision criterion, or criterion boundary, varies is shown in Figure 2; this plot is based on the assumption that the SN and N distributions are Gaussian and of equal variance. A family of curves is shown in this figure; the parameter, d', is an index of signal strength. In particular, d' is defined as the difference between the means of the SN and N distributions normalized to the standard deviation of the N distribution, that is, $d' = \frac{M_{SN} - M_N}{\sigma_N}$. These curves, showing $P_{SN}(A)$ as a function of $P_N(A)$, are called "detectability curves".

Given a particular signal strength, the observer can, according to the theory, operate at any point on the associated detectability curve. Since $P_{SN}(A)$ and $P_N(A)$ are dependent probabilities (an increase in one of these requiring, or resulting in, a determinate increase in the other), a given operating level will be more or less appropriate for a given set of external conditions and a given purpose. The theory of signal detectability[3] specifies the optimum operating level, or criterion, under each of several definitions of optimum. In each case, the optimum criterion is in a family of criteria defined in terms of likelihood ratio. A criterion in this family is denoted $A(\beta)$; that is, the criterion A contains all observations with likelihood ratio greater than β, and none of those with likelihood ratio less than β. With respect to a given definition of optimum the exact value of β to be used constitutes its solution. The slope of a detectability curve at the point corresponding to the optimum operating level is equal to this value of β.

Definitions of Optimum Criteria. Within the theory of signal detectability, several definitions of optimum are advanced, along with their respective solutions. These may be listed as follows:

1) The Weighted-Combination Criterion - the criterion that maximizes $P_{SN}(A) - w\,P_N(A)$. Solution: $A(\beta)$ where $\beta = w$.

2) The Ideal Criterion - the criterion that minimizes total error. Solution: $A(\beta)$ where $\beta = \frac{P(N)}{P(SN)}$, the ratio of a priori probabilities.

3) The Expected-Value Criterion - the criterion that maximizes the total expected value where the individual values are:

$V_{SN\cdot A}$ = the value of a detection

$V_{N\cdot CA}$ = the value of a correct rejection

$K_{SN\cdot CA}$ = the cost of a miss

$K_{N\cdot A}$ = the cost of a false alarm

Solution: $A(\beta)$ where $\beta =$

$$\frac{P(N)}{P(SN)} \cdot \frac{V_{N\cdot CA} + K_{N\cdot A}}{V_{SN\cdot A} + K_{SN\cdot CA}} .$$

4) The Neyman-Pearson Criterion - the criterion that maximizes $P_{SN}(A)$ while $P_N(A) \leq k$. Solution: $A(\beta)$ where $P_N[A(\beta)] = k$.

5) The Information Criterion - the criterion that maximizes the reduction in uncertainty, in the Shannon sense ([5]), as to whether or not a signal was sent. Solution: $A(\beta)$ where

$$\beta = \frac{P(N)}{P(SN)} \cdot \frac{\log P_{CA(\beta)}(N) - \log P_{A(\beta)}(N)}{\log P_{A(\beta)}(SN) - \log P_{CA(\beta)}(SN)} .$$

Conclusions Drawn from Previous Experiments

From the results of previous experiments, it may be stated that the observation variable, x, is continuous; the observer can distinguish among values of x well into the noise ([6,8,11,12,13]). Previous experiments have also shown the observers to be capable of operating in accordance with certain of the optimum criteria listed. In particular, the observer can maximize the expected value of the decision ([6,8,11,12,13]) and he can act as a Neyman-Pearson Observer ([8]). It is likely that the observer can act in accordance with Weighted-Combination and Ideal Criteria, since the Expected-Value Criterion and the Ideal Criterion are special cases of the abstract Weighted-Combination Criterion. The observer can also adapt to another definition of optimum advanced by the theory of signal detectability, one relative to reporting a posteriori probability ([8]); in this case no criterion is assumed, rather the best estimate is made of the probability that the observation arose from signal-plus-noise: $P_x(SN) = \ell(x)\,P(SN) / \ell(x)\,P(SN) + P(N)$. Certain conclusions drawn from previous experiments are more conveniently discussed below.

Theory of Testing Multiple Hypotheses

Given the demonstration that human observers are capable of behavior very nearly optimum in the two-hypothesis task, the present experiments were designed to assess their ability to behave optimally in more complex tasks.

In both of the experiments reported below, the observers are required to decide among three hypotheses. In one case, the observer chooses, on each trial, on the basis of a single observation interval, among the hypotheses: Signal One occurred, Signal Two occurred, Noise Alone was present. In the other case, three hypotheses concerning the time of signal occurrence are tested on each trial: the signal occurred in the first observation interval, in the second observation interval, in the third observation interval.

In these three-hypothesis tests, as in the two-hypothesis tests described above, a criterion approach is required of the observer. He must assume criteria such that each possible value of the observation variable, x, leads to acceptance of one of the given hypotheses. In both experiments, the truth of each of the relevant hypotheses is assigned an a priori probability, and values are assigned to the possible outcomes of the various decisions. This assignment permits definition of the expected value of a decision, and specifies as the appropriate optimum decision rule the one that maximizes the expected value of a decision.

In the experiments, the probabilities and values are varied from one group of trials to another. For each group of trials involving constant values and probabilities, the criteria assumed by the observer are compared with the criteria specified as optimum within the theory. For simplicity of analysis, in the present experiments, positive values were assigned to correct decisions and a value equal to zero was assigned to each incorrect decision.

By definition, the expected value of a decision is the product of probability and desirability of an outcome summed over the possible outcomes. The probability of an outcome in this case is the probability of the joint event involving the signal presented and the signal hypothesis accepted, $P(i\cdot Aj)$. Since $P(i\cdot Aj) = P(i)P_i(Aj)$, for the three-hypothesis test the expected value is defined by

$$EV = \sum_{i=1}^{3} P(i)\, V_i P_i(A_i) \qquad (1)$$

where $P(i)$ is the a priori probability of occurrence of the ith signal, V_i is the value of correctly identifying the occurrence of the ith

signal, and $P_i(A_i)$ is the conditional probability of accepting the occurrence of the ith signal when it occurred. In terms of the density functions

$$EV = \sum_{i=1}^{3} P(i)V_i \int_{A_i} f_i(x)dx \qquad (2)$$

or

$$EV = \sum_{i=1}^{3} \int_{A_i} P(i)V_i f_i(x)dx. \qquad (3)$$

Hence, the optimum decision performance requires that the observation, x, be regarded as belonging to the criterion A_i, (i.e., $x \epsilon A_i$), if and only if

$$P(i)V_i f_i(x) \geq P(j)V_j f_j(x) \qquad (4)$$

for all j.

The development immediately above applies to both experiments. In the first section below, the theory is developed for the case of one of two signals or noise alone, (hereafter referred to as the "detection-and-recognition task") and in the second section below, for the case of one signal in one of three observation intervals (to be referred to as the "forced-choice-in-time task").

The Specification of the Optimum Decision Criteria for the Detection-and-Recognition Task. This section presents the method of deriving the specification of optimum performance to which the observer's performance is compared in the detection-and-recognition task.

Let $f_1(x) = f_{S_1N}(x)$, $f_2(x) = f_{S_2N}(x)$, and

$f_3(x) = f_N(x)$, and call $A_3 = B$.

Then

$$x \epsilon B \Leftrightarrow P(N)\ V_N f_N(x) \geq P(S_j)V_{S_j} f_{S_jN}(x),\ j = 1,2, \qquad (5)$$

that is, x is in B if and only if

$$\frac{P(N)\ V_N}{P(S_j)V_{S_j}} \geq \ell_j(x) = \frac{f_{S_jN}(x)}{f_N(x)},\ j = 1,2. \qquad (6)$$

Similarly,

$$x \epsilon A_2 \Leftrightarrow P(S_2)V_{S_2} f_{S_2N}(x) \geq P(N)V_N f_N(x) \qquad (7)$$

$$\geq P(S_1)V_{S_1} f_{S_1N}(x), \qquad (8)$$

that is,

$$x \epsilon A_2 \Leftrightarrow \ell_2(x) \geq \frac{P(N)V_N}{P(S_2)V_{S_2}} \qquad (9)$$

$$\geq \frac{P(S_1)V_{S_1}}{P(S_2)V_{S_2}}\ \ell_1(x), \qquad (10)$$

and

$$x \epsilon A_1 \Leftrightarrow \ell_1(x) \geq \frac{P(N)V_N}{P(S_1)V_{S_1}} \qquad (11)$$

$$\geq \frac{P(S_2)V_{S_2}}{P(S_1)V_{S_1}}\ \ell_2(x). \qquad (12)$$

In the detection-and-recognition experiment reported below, S_1, S_2, and N were presented with equal probabilities, that is, $P(S_1N) = P(S_2N) = P(N) = \frac{1}{3}$. The optimum criteria varied as a function of changes in the value or payoff matrix. Thus, the optimum decision criteria may be specified as follows:

$$x \notin B \Leftrightarrow \ell_j(x) \geq \frac{V_N}{V_{S_j}}, \text{ for either } j, \qquad (13)$$

and

$$x \epsilon A_1 \Leftrightarrow x \notin B \text{ and } \ell_1(x) \geq \frac{V_{S_2}}{V_{S_1}}\ \ell_2(x). \qquad (14)$$

Equations 13 and 14 permit the assignment of optimum frequencies of response in a 3 x 3 table where the columns represent the true hypothesis and the rows represent the hypothesis accepted, providing the relevant detectability indices and values are known. Just as a decision between two hypotheses can always be represented on a single axis with a normalized Gaussian distribution under noise alone, the decision among three hypotheses can be represented on a plane (two axes) with a normalized Gaussian distribution under noise alone. Consider the model represented in Figure 3. If the signal-plus-noise distributions are merely translations of the noise distribution on the two axes, then the three hypotheses can be represented as three points on a plane in which each distribution is Gaussian in every direction with unit variance. The criterion boundaries are straight since the method of assigning values to the various outcomes of a decision leaves only a single density function in the integral to be maximized (in Equation 4), not linear combinations of density functions. The prediction of the response table then depends upon the distance of the boundaries from the points. Said another way, $\ell_1(x)$ is constant on any line perpendicular to the line $\overline{NS_1}$, and $\ell_2(x)$ is constant on any line perpendicular to the line $\overline{NS_2}$. The ratio of likelihood ratios, $\ell_1(x)$ and $\ell_2(x)$, is also constant along any line perpendicular to the line $\overline{S_1S_2}$. Because the equalities of (13) and (14) hold on optimum boundaries, the three boundaries meet at a point. A knowledge of

the detectability index d', for each pair of hypotheses, that is, of the lengths of the sides of the triangle, and a knowledge of V_{S_1} and V_{S_2}, permits the specification of the optimum response table.*

The method actually used to determine the optimum response probabilities entailed drawing the boundaries on double probability paper ruled in 1% steps on each axis. Each rectangle represents a probability of .01% for a two-dimensional Gaussian distribution centered at the center of the paper. Three graphs are drawn for each observer and each set of values, each graph assuming the truth of one of the three hypotheses. The rectangles within the various boundaries on each graph are counted; thus each graph yields a column of numbers in a 3 x 3 table. This process provides an approximation good to better than 1%. This procedure is described in more detail in the appendix.

The Specification of the Optimum Decision Criteria for the Forced-Choice-in-Time Task. As indicated above, the optimum decision function for the case where multiple hypotheses exist is acceptance of the ith signal when and only when

$$P(i)V_i f_i(x) \geq P(j)V_j f_j(x) \qquad (4)$$

for all j.

In the routine use of the forced-choice-in-time technique, $P(i) = P(j)$ and $V_i = V_j$ for all i and j; in this case only the density function $f_i(x)$ need be considered. In this case, optimum behavior requires that the observation interval yielding the greatest value of $f_1(x)$, or the greatest value of $\ell(x) = \frac{f_{SN}(x)}{f_N(x)}$, be selected as containing the signal. If the signal distribution is a translation of the noise distribution, the probability of a correct response given selection of the largest value of $f_i(x)$, or the probability that the largest value of $f_i(x)$ represents signal-plus-noise, is the probability that one drawing from a normal distribution with mean d' and unit variance is greater than the greater of two drawings from a normal distribution with zero mean and unit variance, where d', it will be recalled, is the difference between the means of the signal-plus-noise and noise distributions normalized to the standard deviation of the noise distribution, $\frac{M_{SN}-M_N}{\sigma_N}$. The probability that a correct answer will result for a given value of d', for the test involving three observation intervals, is given by the equation:

$$P(c) = \int_{-\infty}^{+\infty} \Phi^2(x)\psi(x-d')dx \qquad (15)$$

where $\Phi(x)$ is the area of the noise distribution to the left of x and $\psi(x-d')$ is the ordinate of the signal-plus-noise distribution. P(c) vs. d' for the three-hypothesis task is plotted in Figure 4.

Forced-choice-in-time tests such as the one just described, that is, where $P(i) = P(j)$ and $V_i = V_j$ for all i and j, have been performed prior to the present experiment. The congruence of the estimates of d' from such tests and from several two-hypothesis tests involving variable values and probabilities, in which optimum behavior was demonstrated, indicates that the observer does operate with the optimum decision function in the forced-choice-in-time test. (12) Data from other experiments requiring second choices (8,11), and, in another instance, requiring last choices (13) from the observers in the four-alternative, forced-choice-in-time task, demonstrate that the observation variable is continuous; observers are capable of ordering values of this variable well into the noise. Until the present experiment, however, the first involving unequal values and probabilities, no direct evidence concerning the ability of the human observer to establish optimum criteria in the forced-choice-in-time task was available.

Consider the case represented by the experiment reported below. In this experiment, two conditions were employed. Throughout the first condition $V_1 = V_2 = V_3$; the probabilities were varied from one group of observations to another under the restriction that $P(1) = P(2) \neq P(3)$. Throughout the second condition $P(1) = P(2) = P(3)$; here the values were varied under the restriction that $V_1 = V_2 \neq V_3$. In developing the method of specifying the optimum behavior in forced-choice-in-time tasks, these two conditions may be considered together.

The effects of a priori probabilities and values on the solution for the expected-value definition of optimum in this task are combined in a single parameter, w, where

$$w = \frac{P(3)}{P(1)+P(2)} \cdot \frac{V_3}{V_1}, \quad V_1 = V_2 . \qquad (16)$$

*A previous study reported by Tanner (10) made use of this model of a signal plane where the signal distributions are simple translations of the distribution of noise alone. In that study, the hypotheses were treated pairwise, and the "angle" between the signals, $\arccos\left\{\frac{(d_1')^2 + (d_2')^2 - (d_{12}')^2}{2d_1'd_2'}\right\}$, was studied as a function of frequency separation and duration. The choice of signals — 900 cps, 1000 cps, $\frac{1}{10}$ sec. duration — used in the detection-and-recognition experiment reported below was based on a desire to have a separation large enough to yield a signal angle of almost 90°.

Note that w is defined similarly to the critical value of likelihood ratio, β, for the two-hypothesis task as discussed above.

Recalling that the optimum decision function is to select the greatest value of $P(i)\ V_i f_i(x)$, we may regard equivalently

$$\begin{pmatrix} f_1(x) \\ f_2(x) \\ 2w\ f_3(x) \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \ell_1(x) \\ \ell_2(x) \\ 2w\ \ell_3(x) \end{pmatrix}$$

$$\text{or} \quad \begin{pmatrix} \log_e \ell_1(x) \\ \log_e \ell_2(x) \\ \log_e \ell_3(x) + \log_e 2w \end{pmatrix}$$

as representing the observations associated with the three observation intervals. Now, from Ref.3, Part II, Sections 4.7 and 4.9, if the logarithm of likelihood ratio is normally distributed in noise alone, then the mean is $\frac{-(d')^2}{2}$ and the variance is d'; in signal-plus-noise the mean is $\frac{+(d')^2}{2}$ and the variance is d'. Using the notation H_i: (m, σ) to indicate a normal distribution with mean m and standard deviation σ,

$$\log_e \ell \text{ is in N:} \quad \left(-\frac{(d')^2}{2},\ d'\right)$$
$$\text{and in SN:} \quad \left(+\frac{(d')^2}{2},\ d'\right) \tag{17}$$

and therefore,

$$\frac{\log_e \ell + \frac{(d')^2}{2}}{d'} \text{ is in N: } (0,1)$$
$$\text{and in SN: } \left(\frac{(d')^2}{d'},\ 1\right) = (d',1). \tag{18}$$

Therefore, a convenient form for the decision components is the following:

$$d_i(x) = \frac{1}{d'}\left[\log_e \ell_i(x) + \frac{(d')^2}{2}\right] \quad i = 1,2 \tag{19}$$

$$d_3(x) = \frac{1}{d'}\left[\log_e \ell_3(x) + \frac{(d')^2}{2} + \log_e 2w\right] \tag{20}$$

and if we introduce the parameter

$$k = \frac{1}{d'} \log_e 2w \tag{21}$$

then

$$d_3(x) = \frac{1}{d'}\left[\log_e \ell_3(x) + \frac{(d')^2}{2}\right] + k. \tag{22}$$

That is, three decision components may be considered: $d_1(x)$ and $d_2(x)$ are (0,1) in N and (d',1) in SN, $d_3(x)$ is (k,1) in N and (k+d',1) in SN. Optimum performance requires choosing the interval with the largest associated $d_i(x)$.

Now the probability that $d_3(x)$ is greater than $d_1(x)$ and $d_2(x)$ when the signal is presented in the third interval is the probability that one drawing from the normal distribution (d'+k,1) is greater than two drawings from the normal distribution (0,1). This probability may be determined from Figure 4, since this probability is the same as the probability of a correct response in a three-hypothesis test with equal values and equal probabilities if d' were k larger. The probability that $d_2(x)$ is greater than $d_1(x)$ and $d_3(x)$ when the signal occurs in interval two is the probability that one drawing from the normal distribution (d,1) is greater than one drawing from the distribution (k,1) and one drawing from the distribution (0,1). By symmetry, this is also the probability that $d_1(x)$ is greater than $d_2(x)$ and $d_3(x)$ when the signal occurs in interval one. These three probabilities, for a given d' and a given w, may be determined from the plot of Figure 5. In this figure the ordinate is $P_1(A_1) = P_2(A_2)$, the abscissa is $1-P_3(A_3) = P_3(A_1 \cup A_2)$, and d' and w are parameters. As in the case of the two-hypothesis tests, the optimum operating level is that point on a given detectability curve where its slope is w.

The probability that $d_1(x)$ is greater than $d_2(x)$ when the signal occurs in the second interval (or, by symmetry, the probability that $d_2(x)$ is greater than $d_1(x)$ when the signal occurs in the first interval) may also be computed; this is the probability that one drawing from the normal distribution (0,1) is greater than one from the distribution (d',1) and one from the distribution (k,1). In Figure 6, $P_1(A_1) = P_2(A_2)$ is plotted against $P_1(A_2) = P_2(A_1)$ with d' and w as parameters.

The development just given permits the assignment, for any d' and any w, of optimum frequencies of response in a 3 x 3 table where the columns represent the true hypothesis and the rows represent the hypothesis accepted. Thus, as w, or P(3) relative to P(1) and P(2), or V_3 relative to $V_1 = V_2$, is varied, it is possible to compare the observer's actual performance and the optimum performance for an observer characterized

by his d'.*

The actual computation of optimum response frequencies is conveniently based on a geometrical model similar to the one described above for the detection-and-recognition task. Consider a three-dimensional space, with the three hypotheses represented by points along the axes at distance d'. This space is depicted in Figure 7. Since noise alone is never presented in the forced-choice-in-time task, the present interest lies in the sloping plane of the triangle drawn with dotted lines; this is the plane of Fig. 8. As in the case of the detection-and-recognition task, boundaries are drawn on double probability paper, and the rectangles are counted. (See the appendix).

A Contrasting Prediction of Decision-Making Theory and Threshold Theory for Three-Hypothesis Tests.

According to conventional sensory theory, there exists a lower bound on the boundary of a criterion (a "threshold") at roughly 3 sigma up from the mean of the noise distribution, that is, at a point on the observation continuum that is rarely exceeded by noise alone. Below this point, distinction among different values of the observation variable is assumed to be impossible.

Unlike some earlier papers in this series (8,11,12)whose primary objectives were a description of decision-making theory and threshold theory, the derivation of contrasting predictions from two theories, and the presentation of data relevant to these predictions, the present paper is based on an assumed applicability of the decision-making theory, and is concerned primarily with extending this theory. It seems worthwhile to point out, however, the general nature of the divergence between the two theories with respect to the three-hypothesis tasks considered here. This is conveniently accomplished by presenting a single illustrative example.

Consider the forced-choice-in-time task where one of the three signal hypothoses is a priori more probable or more valuable than the other two, that is, where the signal appears in one of the three intervals with probability greater than one-third and with equal probabilities in the other two intervals, or where a correct choice when the signal occurs in the special interval pays more than a correct choice when the signal occurs in one of the other intervals. Assume, as in the case of the task described above, that the special signal hypothesis, or interval, is the third one. This task falls in the class of tasks characterized by a $k>0$, or a $w>.50$.

According to the threshold theory, if the observation made in any interval is above threshold, this almost certainly indicates that the signal was in this interval. If this happens p_o of the time for a given strength of signal, then the diagonal entries in the conditional probability table will be at least p_o. If no interval yields an observation above threshold, then the observer randomly selects one of the alternatives. The conditional table, in this case, contains two degrees of freedom. In the

		Presented		
		1	2	3
Accepted	1	$p_o+\alpha_1(1-p_o)$	$\alpha_1(1-p_o)$	$\alpha_1(1-p_o)$
	2	$\alpha_2(1-p_o)$	$p_o+\alpha_2(1-p_o)$	$\alpha_2(1-p_o)$
	3	$\alpha_3(1-p_o)$	$\alpha_3(1-p_o)$	$p_o+\alpha_3(1-p_o)$

table, the α_i are non-negative and add to one. If the observer has control of the guessing probabilities α_i, he should concentrate the guessing on the most profitable alternative. The formula for the expected value becomes

$$EV = p_o \sum_{i=1}^{3} P(i)V_i + (1-p_o) \sum_{i=1}^{3} P(i)V_i\alpha_i. \quad (23)$$

In the case in question where $k>0$ $(w>.5)$, he would do best to set $\alpha_3=1$; then the expected table contains four zeros. For the situation

		Presented		
		1	2	3
Accepted	1	p_o	0	0
	2	0	p_o	0
	3	$1-p_o$	$1-p_o$	1

* A similar approach to specifying optimum behavior in the detection-and recognition task, that is, in terms of drawing from normal distributions, might have been taken in the previous section. However, unless the two signals are equally detectable and orthogonal, unless the decision components are distributed independently of one another, this mode of anaylsis involves a correlation term; in this case the analysis is too lengthy to be reported here. It may be seen in the appendix that the requirements of equally detectable and orthogonal signals were not always satisfied.

when $k < 0$ $(w < .5)$, it would be profitable to set $\alpha_3 = 0$; then the expected table contains two zeros. By contrast, the decision-making theory leads to a prediction of positive entries in each cell of the expected table.

The Experiments

In the first part of this section, the apparatus used in the experiments is briefly described. Following this, the two experiments are described and their results presented.

Apparatus

The experiments are essentially conducted by an apparatus called N. P. Psytar. The name of the apparatus is a condensation of the phrase Noise Programmed Psychophysical Tester and Recorder; the details of its construction have been reported elsewhere [4]. Prior to an experimental session, the randomization constants characterizing the experiment, and the physical constants characterizing the signals and noise employed, are effected by knob settings on the machine's face. The machine then programs the actual experiment, that is, determines whether or not a particular signal is presented in a particular observation interval, by sampling the output of a random noise generator; if the voltage of the sample taken during the observation interval exceeds the preset level, a thyratron is triggered and a particular signal is presented.

The observer is informed of the progress of the experiment by a set of three lights. The first light is a warning light; when it flashes a trial cycle begins. The second light flashes in coincidence with the observation interval, or intervals, the number of intervals depending upon the nature of the experiment. The third light indicates that the observer should make a selection by pushing one of the answer buttons available to him. After his selection is recorded by the apparatus, he is informed of its correctness by a second panel of lights which indicates the selection made by the machine.

The signal sources used in the experiments were Hewlett-Packard audio-oscillators 200 AB and 200 I. The masking-noise source was a General Radio 1390A random noise generator, providing a white, Gaussian noise up to 20 kc/s. In the first experiment, the two signals were a 900 cps tone and 1000 cps tone, each of 0.1 second duration. The signal used in the second experiment was a 1000 cps tone of 0.1 second duration.

The signal from the audio-oscillator is fed into a gate circuit; the output of the gate circuit is fed into an adder. The gated signal contains an integral number of cycles, beginning at zero voltage. This is fed to a Williamson amplifier, and then to the observers' PDR-8 earphones. The noise is fed from the noise generator to the adder, to the Williamson amplifier, and then to the earphones. All measurements are made at the input to the earphones, using a Hewlett-Packard 400 B average reading voltmeter calibrated in RMS voltage. Throughout the experiments reported below, the signal strength and noise level were constant; $\frac{2E}{N_o}$ (twice the signal energy divided by the noise power per unit bandwidth) was equal to 13.5 at the earphone input.

The Detection-and Recognition Experiment

Procedure. In the detection-and recognition experiment, five groups of observations, each group containing 400 observations, were obtained. The five groups were distinguished by the values assigned to the observer's correct acceptance of Signal 1 (1000 cps), Signal 2 (900 cps), or Noise alone, respectively. These value clusters in the order presented experimentally, were .6, .2, 1; 1, 1, 1; 2, 3, 1; 1.5. 2, 1; .8, .6, 1. Throughout this experiment, the probability of occurrence of each of the three alternatives was fixed at one-third.

The values listed in the preceding paragraph are listed there in the form in which they were presented to the observers. This form was chosen to make clear the proportionate value of correctly accepting each of the alternatives. As a basis for making actual payoffs to the observers, the various values were equated with fractions of a cent such that the payoff approximated 50 cents per hour throughout the different value conditions. The values employed were selected to yield an adequate sampling of the range within which none of the expected response frequencies drop below a minimum value required by certain statistical tests performed.

Two college students, one a music major and the other an English major, served as observers in this experiment and also in the second experiment reported below. They observed two hours a day for five days a week; they received a regular hourly rate for their services, as well as the payoff based on their performance. The observers were rather thoroughly experienced before the experiments reported here were begun. Prior to these experiments, they had taken part in an extensive investigation of the influence of the amount of prior information on the approach to the optimum decision criteria in a two-hypothesis test, which involved over 10,000 observations [8].

The Data. The raw data, in terms of the response frequencies obtained, are presented in the 3 x 3 tables of Table I. The numbers within parentheses in each cell are the optimum response frequencies, the frequencies predicted from the theory.

Values Assigned: $S_1N = .6$, $S_2N = .2$, $N = 1$

Observer 1

Accepted \ Presented	S_1N	S_2N	N	
S_1N	91 (90.6)	28 (26.0)	29 (19.0)	
S_2N	18 (3.8)	72 (38.3)		
N	20 (34.6)	38 (72.7)	105 (115.0)	
	129	138	134	401

Observer 2

Accepted \ Presented	S_1N	S_2N	N	
S_1N	93 (81.4)	25 (16.2)	16 (23.1)	
S_2N	3 (2)	64 (42.8)		
N	73 (45.0)	49 (79.0)	118 (110.9)	
	129	138	134	401

Values Assigned: $S_1N = 1$, $S_2N = 1$, $N = 1$

Observer 1

Accepted	S_1N	S_2N	N	
S_1N	95 (98.6)	24 (12.1)	56 (53.6)	
S_2N	18 (9.7)	93 (96.3)		
N	13 (17.7)	17 (25.9)	85 (87.4)	
	126	134	141	401

Observer 2

Accepted	S_1N	S_2N	N	
S_1N	61 (63.7)	14 (5.6)	39 (42.0)	
S_2N	27 (18.1)	91 (98.2)		
N	38 (44.2)	29 (30.2)	102 (99.0)	
	126	134	141	401

Values Assigned: $S_1N = 2$, $S_2N = 3$, $N = 1$

Observer 1

Accepted	S_1N	S_2N	N	
S_1N	97 (95.9)	14 (17.2)	78 (75.4)	
S_2N	34 (30.1)	120 (120.1)		
N	2 (6.4)	11 (7.7)	44 (46.6)	
	133	145	122	400

Observer 2

Accepted	S_1N	S_2N	N	
S_1N	71 (72.0)	13 (3.9)	106 (100.6)	
S_2N	59 (54.6)	129 (139.2)		
N	3 (6.4)	3 (1.9)	16 (21.4)	
	133	145	122	400

Values Assigned: $S_1N = 1.5$, $S_2N = 2$, $N = 1$

Observer 1

Accepted	S_1N	S_2N	N	
S_1N	92 (92.9)	19 (27.8)	80 (68.0)	
S_2N	33 (24.9)	102 (97.6)		
N	3 (10.2)	17 (12.6)	54 (66.0)	
	128	138	134	400

Observer 2

Accepted	S_1N	S_2N	N	
S_1N	86 (78.5)	11 (17.5)	76 (88.0)	
S_2N	33 (35.3)	119 (110.0)		
N	9 (14.2)	8 (10.4)	58 (46.0)	
	128	138	134	400

Values Assigned: $S_1N = .8$, $S_2N = .6$, $N = 1$

Observer 1

Accepted	S_1N	S_2N	N	
S_1N	79 (103.0)	17 (19.6)	33 (33.5)	
S_2N	40 (12.2)	82 (71.6)		
N	22 (25.8)	28 (35.7)	99 (98.5)	
	141	127	132	400

Observer 2

Accepted	S_1N	S_2N	N	
S_1N	95 (80.6)	12 (8.3)	26 (33.3)	
S_2N	9 (12.5)	85 (79.6)		
N	37 (47.9)	30 (39.1)	106 (98.7)	
	141	127	132	400

TABLE I. The Raw Data from the Detection-and-Recognition Experiment

The Correlation Between Predicted and Obtained Response Tables. As discussed above, the optimum criteria are specified by Equations 13 and 14. These equations, along with the value of correctly accepting each alternative, and the detectability index (d') characterizing each pair of alternatives, determine the optimum response frequencies. The values of d' for each observer, for each pair of alternatives, are obtained in separate experiments. Here, the detectability indices, d'_1 and d'_2, for S_1N and S_2N respectively, are obtained individually in three-alternative, forced-choice experiments. The index d'_{12} is obtained in an experiment where either S_1N or S_2N is presented in the single interval of a trial cycle. For observer 1, $d'_1 = 1.77$, $d'_2 = 1.45$ and $d'_{12} = 1.84$; for Observer 2, $d'_1 = 1.45$, $d'_2 = 1.45$ and $d'_{12} = 2.04$. The values of d'_1 and d'_2 listed are based on 400 observations and the value of d'_{12} on 600 observations.

Table II gives the coefficients of rank-order correlation (Spearman's rho) between the individual predicted and obtained response tables of Table I. For samples of this size, a coefficient of .62 has an associated probability of .05 and a coefficient of .79 has an associated probability of .01, under the null hypothesis [2]. Note in Table II that nine of the ten coefficients have an associated probability of less than .01, and the tenth has an associated probability between .05 and .01.

V_{S_1N}	V_{S_2N}	V_N	Observer 1	Observer 2
.6	.2	1	.88	.92
1	1	1	.81	.98
2	3	1	1.00	.95
1.5	2	1	.98	1.00
.8	.6	1	.67	1.00

TABLE II. The Correlation between Predicted and Obtained Response Frequencies in the Detection-and-Recognition Experiment

The Comparison of Predicted and Obtained Payoffs. It is instructive to compare the payoff obtained by the observer with certain theoretical amounts of payoff. One of these, the one of primary interest, is the maximum payoff attainable, given the observer's sensitivity characteristics (d'_1, d'_2 and d'_{12}); a second is the payoff attainable by an infinitely sensitive observer; still a third is the payoff attainable without sensitivity, or with the earphones removed, by simply choosing consistently the most valuable alternative.

Table III presents these various payoffs, for each observer, for each value condition. They are normalized such that the payoff attainable with infinite sensitivity is equal to unity.

	Value Conditions / Payoff basis	.6,.2,1	1,1,1	2,3,1	1.5,2,1	.8,.6,1
Observer 1	$d' = \infty$	1.00	1.00	1.00	1.00	1.00
	predicted	.741	.688	.727	.665	.697
	obtained	.728	.693	.727	.658	.658
	$d' = 0$	.561	.352	.529	.458	.411
Observer 2	$d' = \infty$	1.00	1.00	1.00	1.00	1.00
	predicted	.704	.651	.743	.692	.640
	obtained	.781	.633	.662	.706	.726
	$d' = 0$	.561	.352	.529	.458	.411

TABLE III. A Comparison of Various Payoff' in the Detection-and-Recognition Experiment

Observer \ Value Condition	.6,.2,1	1,1,1	2,3,1	1.5,2,1	.8,.6,1
1	.98	.99	1.00	.99	.94
2	1.11	.97	.89	1.02	1.14

TABLE IV. The Ratio of Obtained to Predicted Payoff in the Detection-and-Recognition Experiment

An overall impression of these data is more easily obtained from their graphical presentation in Figure 9. It may be observed from this figure that the largest discrepancy between prediction and data is approximately 7% for Observer 1 and 18% for Observer 2.

That, in some instances, the payoff obtained is greater than the payoff predicted, is explained by the fact that the estimates of the observer's sensitivity characteristics (which contribute to the determination of the value predicted) are based on separate experiments, performed on different days, and are not extracted from the data from which the value obtained is computed.

An index of the congruence of the predicted and obtained payoffs may be had by taking the ratio of the two. The ratio of obtained to predicted payoff, for each condition, is presented in Table IV.

Actually, the observers may not be performing as near optimum as the data presented in Figure 9 and Table IV make it appear. It is quite possible that the payoff scheme chosen is one which is relatively insensitive to deviations from the optimum criteria, and, hence, that the ratio of obtained to predicted payoff is a somewhat misleading index of the congruence between the optimum criteria and the criteria assumed by the observers. It may be, in other words, that the payoff scheme employed, inadvertently, yields an almost constant payoff over a relatively wide range of criteria in the critical region. Whether or not this is the case is difficult to ascertain directly, since there are two degrees of freedom in the criteria corresponding to a given value cluster, and hence the expected payoff as a function of change in criteria cannot be represented in two dimensions. This difficulty is overcome in the forced-choice-in-time experiment reported below, since in the latter experiment a symmetry exists between two of the signal alternatives; this means that the criteria corresponding to a given value matrix and a given set of a priori probabilities may be assumed to possess a single degree of freedom, thus permitting a simple display of expected payoff as a function of change in criteria. This problem is taken up again in the context of the forced-choice-in-time experiment discussed below.

An Analysis in Terms of Chi-Square. Several Chi-square tests were applied to the response tables in the detection-and-recognition experiment. The complete numerical results are presented in a forthcoming technical report[7]. One set of Chi-square tests tested the obtained response tables against those predicted by using the sensitivity characteristics (d'_1, d'_2, d'_{12}) obtained in the control experiments and the optimum criteria. In all cases, a very significant difference exists. For certain cases, the boundaries were then shifted to lower the Chi-square, still using the equal-variance, Gaussian model and the sensitivity characteristics estimated from the control experiments. This procedure reduced the values of Chi-square considerably. Because of the cut-and-try method used, and the rather large amount of computation time required to obtain each new set of response frequencies, no program of analysis was conducted. However, enough was done to say that the primary contribution to the Chi-square resulted from the assumption that the criteria assumed by the observers were actually the optimum criteria, small changes in the location of the criteria ($\frac{1}{10}$ of a standard deviation) causing large changes in Chi-square and practically no change in expected payoff.

The Forced-Choice-in-Time Experiment

Procedure. In the forced-choice-in-time experiment, two main conditions were employed. In the first, the value associated with correctly accepting each of the three alternatives was constant ($V_1 = V_2 = V_3 = 1$), whereas the a priori probabilities of occurrence of each of the alternatives were changed from one subcondition to another, from $P_1 = P_2 = .445$ and $P_3 = .11$ to $P_1 = P_2 = .167$ and $P_3 = .666$. In the second main condition, the situation was the reverse; the probabilities were constant ($P_1 = P_2 = P_3 = .333$), and the values were changed from $V_1 = V_2 = 1$ and $V_3 = 4$ to $V_1 = V_2 = 4$ and $V_3 = 1$. In the first main condition, the a priori probabilities were selected to yield w's of 2.00 and .125 (see Equation 16); in the second main condition, two different sets of values were also selected to yield w's of 2.00 and .125. In each of the four subconditions, 400 observations were obtained.

The Data. As in the case of the preceding experiment, the raw data are presented in 3 x 3 tables, in Table V. Again the optimum or predicted response frequencies are listed in parentheses. The data in the first four individual tables were obtained with $V_1 = V_2 = V_3 = 1$; the data in the last four individual tables were obtained with

$P_1 = P_2 = .445$, $P_3 = .11$

$w = .125$

Presented

Accepted	1	2	3	
1	300 (313.0)		8 (14.0)	
2	46 (38.3)			
3		18 (12.7)	28 (22.0)	
	182	182	36	400
		364		

Observer 1

Presented

Accepted	1	2	3	
1	270 (291.0)		11 (21.0)	
2	69 (58.2)			
3		25 (14.8)	25 (15.0)	
	182	182	36	400
		364		

Observer 2

$P_1 = P_2 = .167$, $P_3 = .666$

$w = 2.00$

Presented

Accepted	1	2	3	
1	78 (73.0)		20 (12.0)	
2	4 (6.7)			
3		17 (19.3)	181 (189.0)	
	46	53	201	300
		99		

Observer 1

Presented

Accepted	1	2	3	
1	62 (66.0)		11 (14.0)	
2	2 (7.5)			
3		35 (25.5)	190 (187.0)	
	46	53	201	300
		99		

Observer 2

$V_1 = V_2 = 1$, $V_3 = 4$

$w = .125$

Presented

Accepted	1	2	3	
1	211 (216.0)		55 (74.0)	
2	43 (43.2)			
3		16 (10.8)	75 (56.0)	
	129	141	130	400
		270		

Observer 1

Presented

Accepted	1	2	3	
1	199 (208.0)		66 (86.0)	
2	60 (51.3)			
3		11 (10.7)	64 (44.0)	
	129	141	130	400
		270		

Observer 2

$V_1 = V_2 = 4$, $V_3 = 1$

$w = 2.00$

Presented

Accepted	1	2	3	
1	123 (127.0)		11 (13.0)	
2	13 (21.0)			
3		64 (91.9)	149 (147.0)	
	120	120	160	400
		240		

Observer 1

Presented

Accepted	1	2	3	
1	90 (115.0)		6 (13.0)	
2	6 (21.6)			
3		144 (103.4)	154 (147.6)	
	120	120	160	400
		240		

Observer 2

TABLE V. The Raw Data from the Forced-Choice-in-Time Experiment

$P_1 = P_2 = P_3 = .333$. These tables are essentially five-fold tables; the respective numbers of correct responses to the first two alternatives are not distinguished as they were in the experiment reported above. It may be noted, relative to the discussion of the threshold theory above, that none of the cell entries is zero.

The Data in Relation to Detectability Curves. The data of Table V, converted to conditional probabilities, are plotted among the detectability curves specified by the theory, in Figure 10. The number labelling a point indicates the observer from which the point was obtained. The letter P following the number indicates that the point was obtained under a condition in which the a priori probabilities were unequal; the letter V indicates a condition in which unequal values distinguished the signal alternatives. (The consistent difference between the estimates of d' obtained when the probabilities were varied and when the values were varied is tentatively associated with a drift in the meter used to set the noise power.) The data points falling to the left of the line w = .50 were obtained under those conditions characterized by a w = 2.00; the points to the right of this line were obtained when a w = .125 was in effect.

For the sake of comparison, the same points are plotted among the detectability curves specified by the threshold theory, in Figure 11. These curves are traced under the assumption that $P_1(A_1) = P_2(A_2)$ is equal to the p_o associated with a given strength signal (the "true" detection probability, the probability that the observation will exceed the threshold) plus a chance factor which equals one-half of $P_3(A_1 \cup A_2)$. The top curve on the graph, $p_o + \frac{1}{2}(1-p_o)$, is a reflection of the bottom curve, p_o, about the axis $P_1(A_1) = P_2(A)_2 = .50$. It defines the upper limit of $P_3(A_1 \cup A_2)$, which is presumably equal to $1-p_o$, and of $P_1(A_1) = P_2(A_2)$, for a given strength of signal.

It is immediately apparent from Figures 10 and 11 that the present experiment was not designed to yield data which might be said to fit one set of curves better than the other. There is, however, one basis on which Figures 10 and 11 may be used in a comparison of the present data with the two theories: given that the observer can control $P_3(A_1 \cup A_2)$, since it shifts appropriately with a change in w, under the threshold theory one would predict that all of the data points would fall either on the line $p_o + \frac{1}{2}(1-p_o)$ or on the line $P_3(A_1 \cup A_2) = 0$.

The Correlation Between Predicted and Obtained Response Tables. The optimum, or predicted, response frequencies for the forced-choice-in-time task are obtained, given the value of d' and w, from the detectability curves of Figures 5 and 6. As just noted, each of the four subconditions yields a point on each of these graphs. The optimum conditional probabilities for a value of d', so determined, and the appropriate value of w, are read directly from the graphs; these conditional probabilities are then converted to response frequencies. Unlike the procedure in the first experiment reported, the optimum response frequencies that are compared with a given set of obtained response frequencies are determined in part by the d' characterizing only that set of obtained response frequencies. In this experiment, in other words, the sensitivity characteristic of the observer is not determined in a separate control experiment and then assumed constant, for purposes of analysis, throughout the experiment.

Table VI presents the coefficients of rank-order correlation between the individual predicted and obtained response tables of Table V. For samples of this size, a coefficient of .80 has an associated probability of .05 and a coefficient of .90 has an associated probability of .01, under the null hypothesis (2).

	Probabilities Variable		Values Variable	
	w=.125	w=2.00	w=.125	w=2.00
Observer 1	.90	.90	.90	1.00
Observer 2	.67	1.00	.90	.87

TABLE VI. The Correlation between Predicted and Obtained Response Frequencies in the Forced-Choice-in-Time Experiment.

The Comparison of Predicted and Obtained Payoffs. As in the case of the previous experiment, the payoff earned by the observer may be compared with the payoff attainable given his demonstrated sensitivity characteristic, and with the payoffs attainable with $d' = \infty$ and $d' = 0$. These amounts are listed in Table VII and presented graphically in Figure 12. The amounts are normalized so that the payoff attainable with infinite sensitivity is equal to 1.00. It may be seen in Figure 12 that the largest discrepancy between prediction and data is approximately 5% for Observer 1 and 8% for Observer 2.

The indices of the approach of the obtained payoff to the predicted payoff, in terms of the ratio of these quantities, are given in Table VIII.

By recourse to plots of the expected payoff vs. $P_3(A_1 \cup A_2)$, it may be stated that, in this experiment, the observers are very probably not performing as strikingly near optimum as indicated by the data displayed in Figure 12 and Table VIII. The plots of expected payoff

		Probabilities Variable		Values Variable	
		w = .125	w = 2.00	w = .125	w = 2.00
Observer 1	d' = ∞	1.00	1.00	1.00	1.00
	predicted	.838	.873	.760	.813
	obtained	.820	.863	.760	.817
	d' = 0	.455	.670	.446	.727
Observer 2	d' = ∞	1.00	1.00	1.00	1.00
	predicted	.765	.843	.724	.799
	obtained	.728	.840	.711	.802
	d' = 0	.455	.670	.446	.727

TABLE VII. A Comparison of Various Payoffs in the Forced-Choice-in-Time Experiment.

	Probabilities Variable		Values Variable	
	w = .125	w = 2.00	w = .125	w = 2.00
Observer 1	.98	.99	1.00	1.01
Observer 2	.96	1.00	.98	1.00

TABLE VIII. The Ratio of Obtained to Predicted Payoff in the Forced-Choice-in-Time Experiment

vs. $P_3(A_1 \cup A_2)$ presented in Figures 13 and 14, for w =.125 and w = 2.00 respectively, show an almost constant payoff to obtain over a considerable range of $P_3(A_1 \cup A_2)$. For w =.125, the payoff is essentially constant from $P_3(A_1 \cup A_2)$ = .40 to .98 for the lowest value of d' obtained, d' = 1.2 (see Figure 10), and from .10 to .98 for the highest value of d' obtained (d' = 1.8). It should also be noted, however, in Figure 10, that the observers are not making full use of this latitude. The discrepancies between the obtained estimate of $P_3(A_1 \cup A_2)$ and the optimum $P_3(A_1 \cup A_2)$ for w =.125 are .18 and .15 for Observer 1 and .26 and .16 for Observer 2. For w = 2.00, the payoff curve is fairly flat from $P_3(A_1 \cup A_2)$ = .01 to .20. The deviationsof obtained from optimum $P_3(A_1 \cup A_2)$ are .04 and .01 for Observer 1 and .07 and .04 for Observer 2. This analysis,

however, does make it clear that one of the more appropriate measures of congruence of obtained and optimum performance may be quite insensitive, unless the payoff scheme is carefully constructed.

Summary and Conclusions

The concern in this paper is limited to a particular type of detection problem, that involving a fixed, finite observation interval. The observation made by the observer is assumed to be mapped onto a small-dimensional space for purposes of decision; the observation made in a two-hypothesis test is mapped onto a one-dimensional space, and the observation in a three-hypothesis test is mapped onto a two-dimensional space. In the theory of optimum detection, such a mapping is a monotone function of the likelihood ratios of the observation. In the application of this theory to the human observer, such a mapping is considered fixed but not necessarily optimum; however, the subdivision of the space into criteria will be related to the likelihood ratios on this reduced observation space. As a matter of convenience, it is assumed that the noise distribution in the reduced observation space is normalized Gaussian; this is not a restriction of generality, but a degree of freedom at the analyst's disposal.

In detectability theory, many related definitions of "optimum" are used. A single one of these, the expected-value definition, is considered in this paper. The observer is informed of the existing a priori probabilities of all events and the value associated with each possible outcome of a decision. His ability to adjust the decision criteria to maximize the payoff is studied. In the experiments reported, the value of an incorrect identification was zero, of a correct response, positive. This scheme was employed for two reasons. The first was to simplify the presentation of the task to the observer. Perhaps more important is that this scheme makes each decision function linear with one log likelihood ratio, and under the condition of a normalized Gaussian noise, the usual distance in the space is proportional to log likelihood ratio in a manner independent of the detectability indices. This simplifies the model and makes computation comparatively simple. A geometric model using likelihood ratios directly as axes will also have straight-line boundaries in all cases of value assignments (see Fig. 1 of Reference 15), but the distributions in such a case make computation difficult.

Two types of the three-hypothesis test are studied. A detection-and-recognition experiment involving the hypotheses Signal One, Signal Two, and Noise Alone, is related to previous work (10) which treated the hypotheses pairwise. In the present paper, the hypotheses were treated pairwise to obtain the configuration of the reduced observation space; this configuration, together with the optimum placement of boundaries of criteria, permits one to make predictions. A forced-choice-in-time experiment reported is the first one to employ an unbalance among the signal hypotheses, thus demanding study of the role of criteria adjustment. In this experiment, one degree of symmetry among the hypotheses was retained in order to determine the sharpness of the optimum, and hence the relative sensitivity of the experiment to the task assigned the observer, that of maximizing the expected value.

Two methods of manual computation of the predicted response frequencies are presented. In these methods, particular emphasis is placed upon the numerical specification of the distance of the optimum boundaries from the hypotheses means as functions of the values, a priori probabilities, and the detectability index; and upon a geometrical model, showing the hypothesis configuration and the placement of the boundaries.

The exact experimental procedure is reviewed. The data from two observers are tabulated in response-frequency tables, together with predictions based on measured sensitivities and optimum criteria. The correlation between prediction and data is shown to be highly significant. The expected value is compared with the payoffs obtained by the observer; the discrepancy is found to be very small percentagewise. Some response-frequency tables were examined using a Chi-square test; this test indicates that the criteria assumed by the observers are definitely not the optimum criteria. However, in view of the fact that the expected value, which constituted the observers' motivation, is not sharply distributed, the Chi-square test may be regarded as overly sensitive.

In a study of this sort, it is appealing to use a probability model that is a reasonable picture of the physical situation represented by the experiment. In the present case, then, an attractive model would be one involving representation of a pulse signal with unknown carrier phase and some uncertainty with respect to starting time and duration, and possibly of frequency-scanning characteristics. Computation of predicted response-frequency tables with asymmetric criteria under such a model is extremely difficult. Studies designed to explore the applicability of this more complex model are being conducted, but, in general, they avoid the problem studied here, that of criteria adjustment, by relying on obvious physical symmetry to force symmetric criteria, and by using only the average correct score as the measurement. For the purposes of criteria-adjustment studies at these levels of noise (d'_i of 1.0 to 2.0), the extremely simple model, in which the signal-plus-noise distribution is a translate of the distribution of noise alone, and the signals are at measurable angle with each other, proves to be an adequate model for prediction of expected payoff and of response-frequency tables to within a few percent.

It may be worthwhile to emphasize that the congruence of predictions and data in a normative study like this one has a dual significance. Whereas the concern in theoretically-oriented

experimental activity lies usually with the matching of theory to observed behavior whatever it may be, in this sort of study an interest in the subject's ability to match the predictions that follow from the theory achieves equivalent status. The present study demonstrates that a simple form of the decision-making model permits prediction of the detection and recognition behavior of the human observer in fairly complex situations and, in addition, indicates the extent to which intelligence may influence a process usually assumed to involve primarily fixed parameters.

References

1. Birdsall, T. G. "The Theory of Signal Detectability" In Quastler, H., (ed) Information Theory in Psychology. Glencoe, Ill.: Free Press, 1955.

2. Kendall, M. G. The Advanced Theory of Statistics. London: J. B. Lippincott Co., 1943.

3. Peterson, W. W. and Birdsall, T. G., "Theory of Signal Detectability," Part I and II, Tech. Rpt. No. 13, Electronic Defense Group, University of Michigan, 1953. Also available in 1954 Symposium on Information Theory, Transactions of the IRE, PGIT-4, September, 1954.

4. Roberts, G. A., "An Automatic Random Programmer," Department of Electrical Engineering, Engineering Research Institute, Electronic Defense Group, University of Michigan.

5. Shannon, C. E. and Weaver, W. The Mathematical Theory of Communications. University of Illinois Press, 1949.

6. Swets, J. A. The Influence of Various Amounts of Prior Information on Decision-Making. Technical Report, Electronic Defense Group, University of Michigan (in preparation).

7. Swets, J. A. and Birdsall, T. G. "Decision-Making in Detection and Recognition Situations Involving Multiple Alternatives. Technical Report, Electronic Defense Group, University of Michigan (in preparation).

8. Swets, J. A., Tanner, W. P., Jr., and Birdsall, T. G., "The Evidence for a Decision-Making Theory of Visual Detection," Tech. Rpt. No. 40, Electronic Defense Group, University of Michigan, 1955.

9. Tanner, W. P., Jr., "On the Design of Psychophysical Experiments" In Quastler, H. (ed) Information Theory in Psychology. Glencoe, Ill.: Free Press, 1955.

10. Tanner, W. P., Jr., "A Theory of Recognition," Technical Report No. 50, Electronic Defense Group, University of Michigan, 1955.

11. Tanner, W. P., Jr., and Swets, J. A. "The Human Use of Information: I. Signal Detection for the Case of the Signal-Known-Exactly. Transactions of the IRE, PGIT-4, September, 1954.

12. Tanner, W. P., Jr., Swets, J. A., "A Decision-Making Theory of Visual Detection," Psychol. Rev., Vol. 61, No. 6, 1954.

13. Tanner, W. P., Jr., Swets, J. A., and Green, D. M., "Some General Properties of the Hearing Mechanism," Tech. Rpt. No. 30, Electronic Defense Group, University of Michigan, 1956.

14. Van Meter, D. and Middleton, D., "Modern Statistical Approaches to Reception in Communication Theory," 1954 Symposium on Information Theory, Transactions of IRE, PGIT-4, September, 1954.

15. Van Meter, D. and Middleton, D., "On Optimum Multiple-Alternative Detection of Signals in Noise," IRE Transactions on Information Theory, Vol. IT-1, September, 1955.

Appendix

In order to predict the 3 x 3 tables, the probabilities of the criteria in a 2-dimensional normal distribution must be computed. This was done by a graphical counting technique. First it is observed that each square on double-probability paper,* ruled in one-percent steps, has a probability of 10^{-4} for a 2-dimensional normal distribution centered at the 50 percent-50 percent intersection, with the proper standard deviation. Therefore a three-hypothesis triangle is drawn to scale with any convenient orientation as in Fig. 15.

The position of the boundaries for a given three-hypothesis triangle is determined from Equations 13 and 14. Then the triangle and boundaries are drawn on three double-probability grids, one for each mean at the center of the paper.

Figures 16 through 18 display these graphs drawn on paper with squares of probability 10^{-2}, for demonstration and not for computation.

This computational method was adopted because only certain specific tables were desired. For the case of equally-difficult, orthogonal signals i.e., for $d'_1 = d'_2 = \sqrt{0.5}\, d_{12}$, tables were computed which allowed interpolation to $\frac{1}{2}$ percent accuracy, and the nine-fold tables could be obtained easily from these. The tables, of course, were derived from computation of double-probability grids.

---

* Codex Graph Sheet No. 42453. The abscissa and ordinate scales are each $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-x^2/2)dx$, where t is a linear scale.

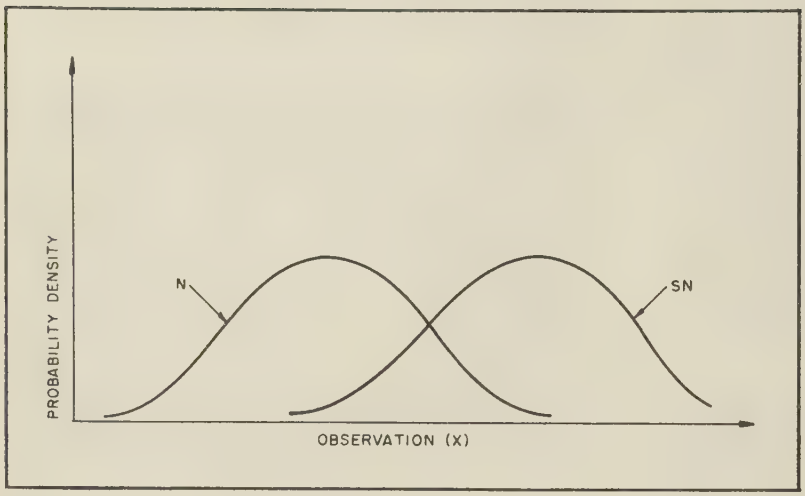


Fig. 1 - The distribution of noise and signal-plus-noise.

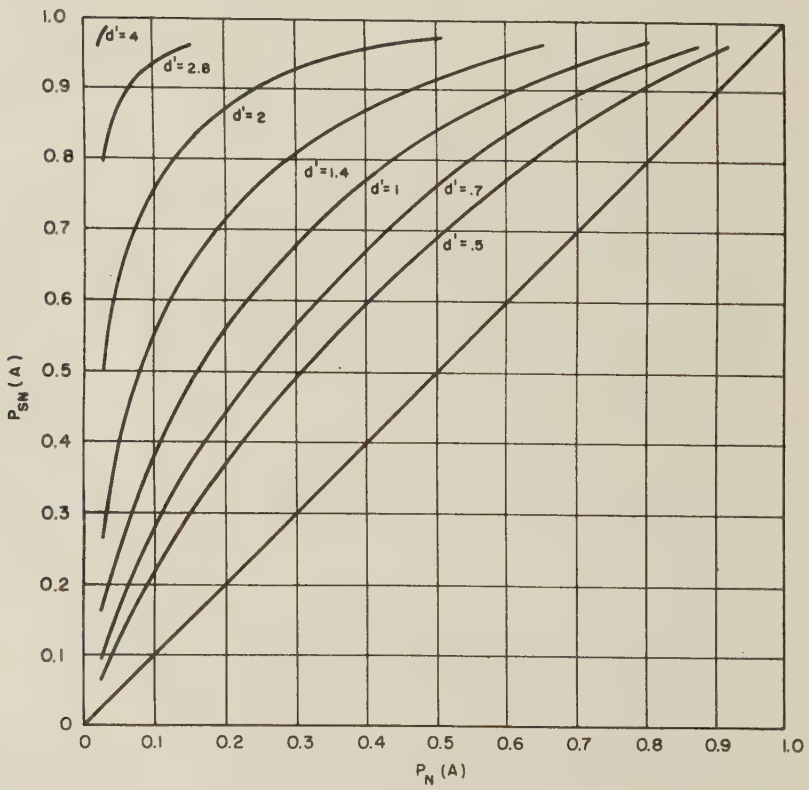


Fig. 2 - $P_{SN}(A)$ vs $P_N(A)$ with d' as the parameter.

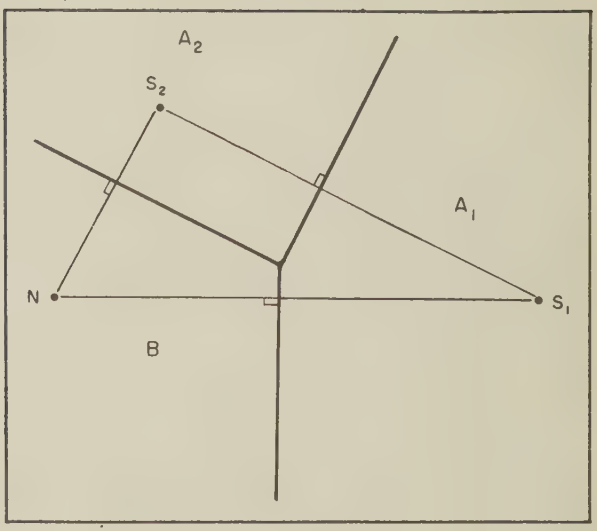


Fig. 3 - A geometrical model of the detection-and-recognition task.

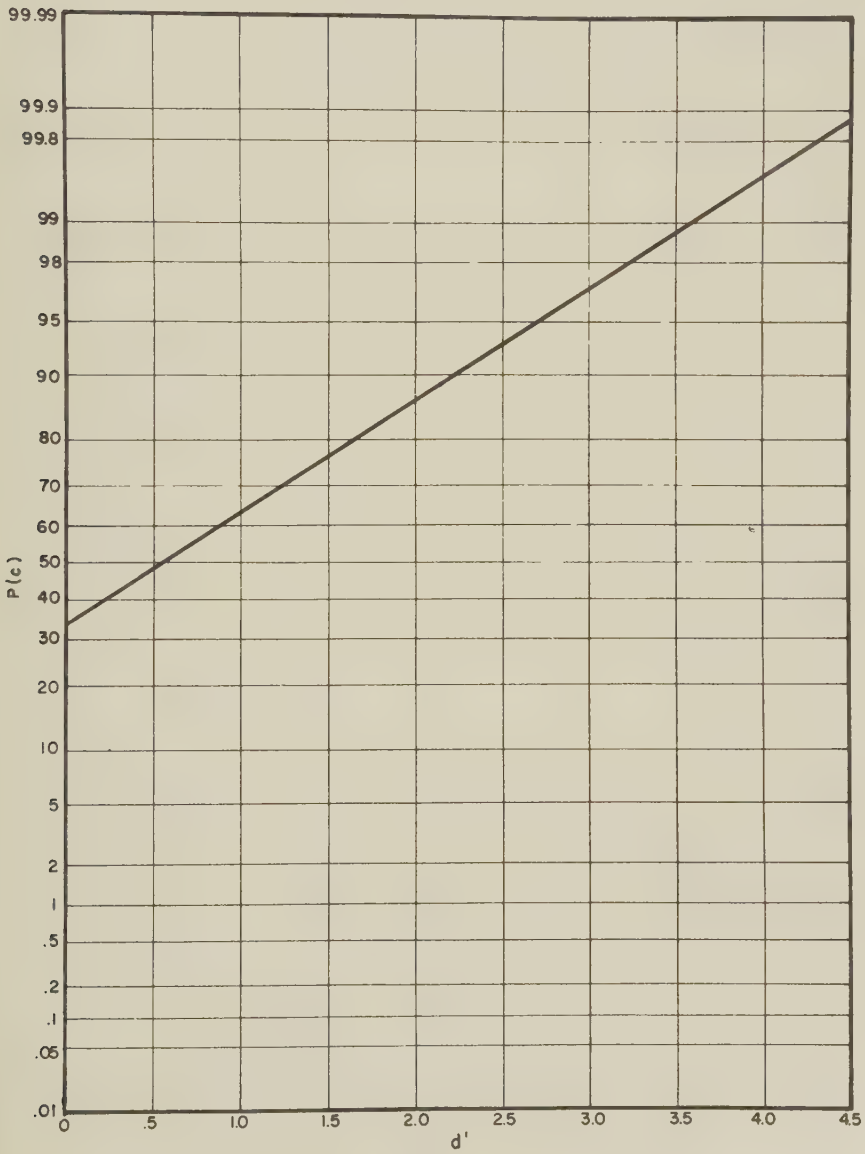


Fig. 4 - P(c) vs d' for the three-alternative, forced-choice task.

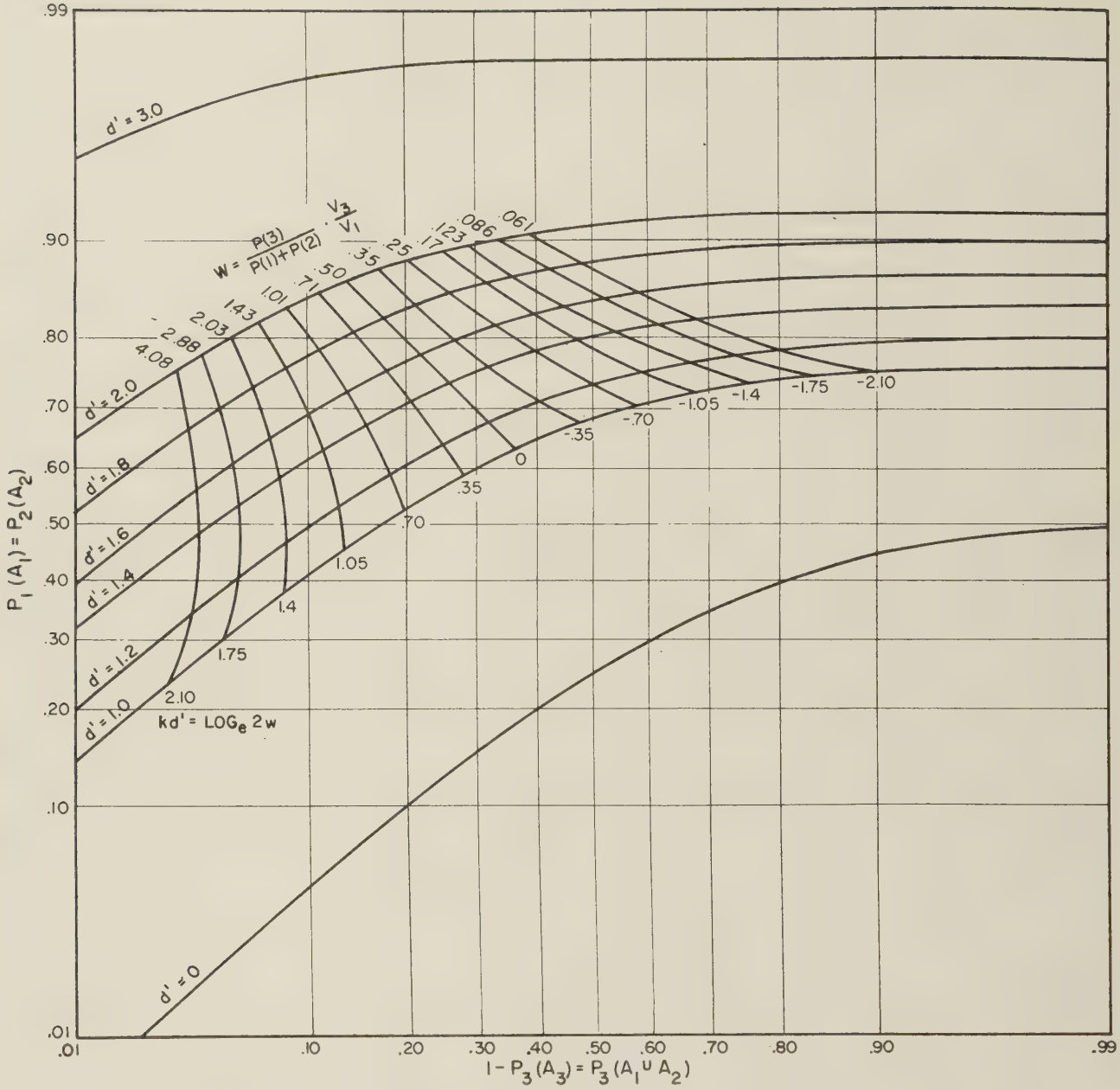


Fig. 5 - $P_1(A_1)$ vs $P_3(A_1 \cup A_2)$ with d' and w as parameters.

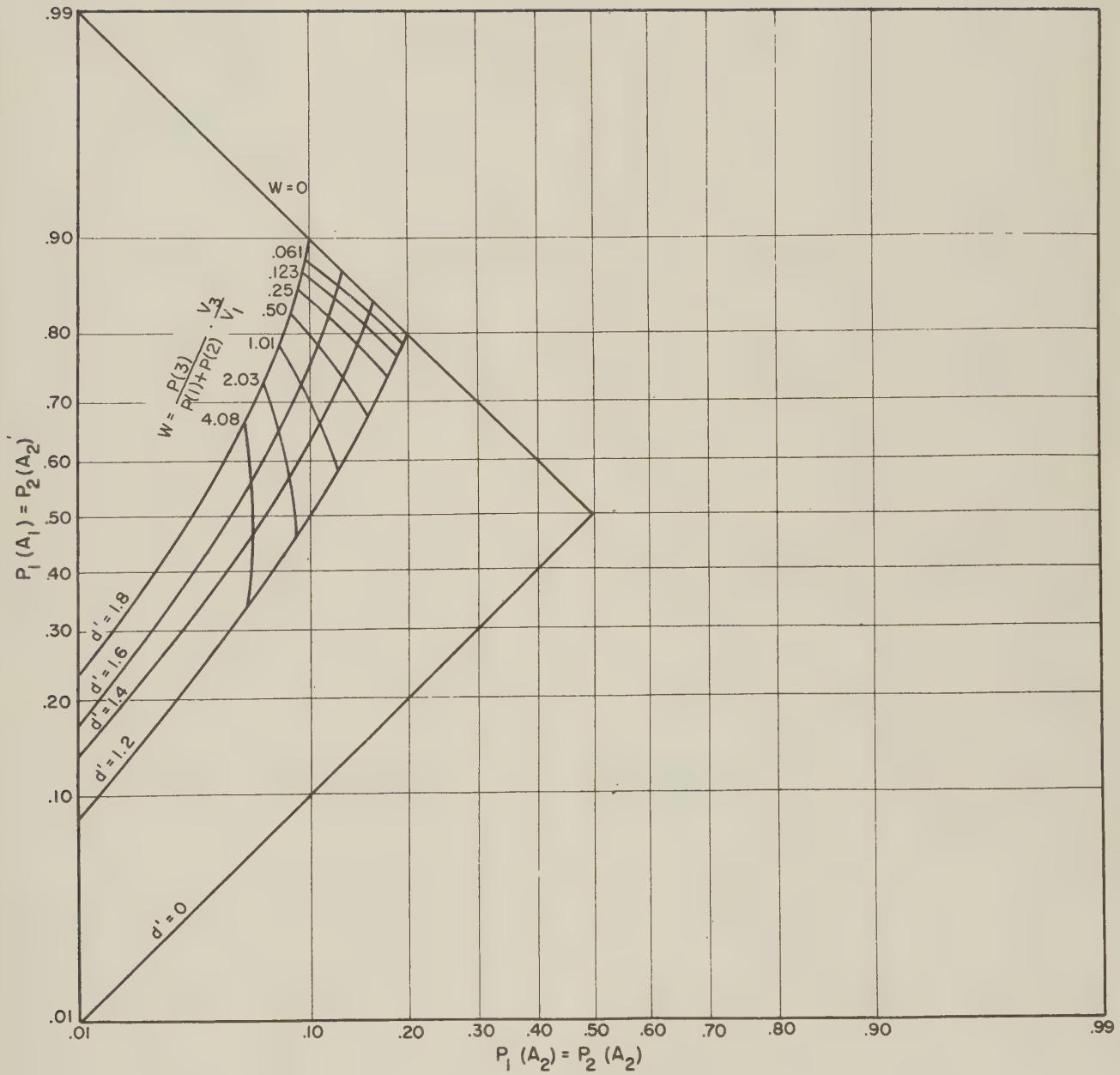


Fig. 6 - $P_1(A_1)$ vs $P_1(A_2)$ with d' and w as parameters.

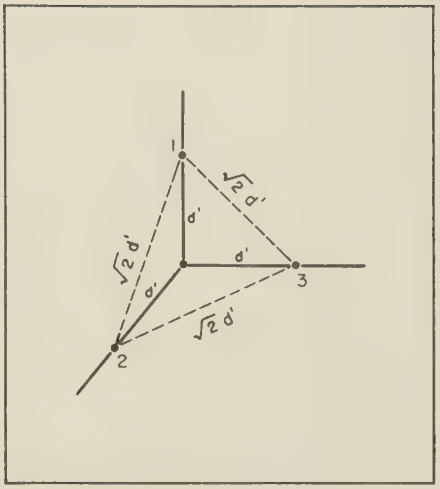


Fig. 7 - A geometrical model of the forced-choice-in-time task, A.

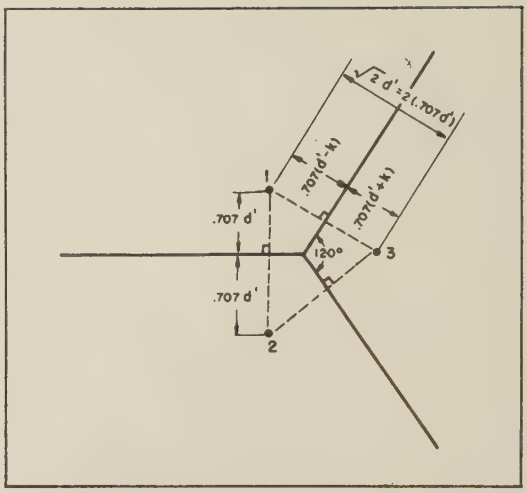


Fig. 8 - A geometrical model of the forced-choice-in-time task, B.

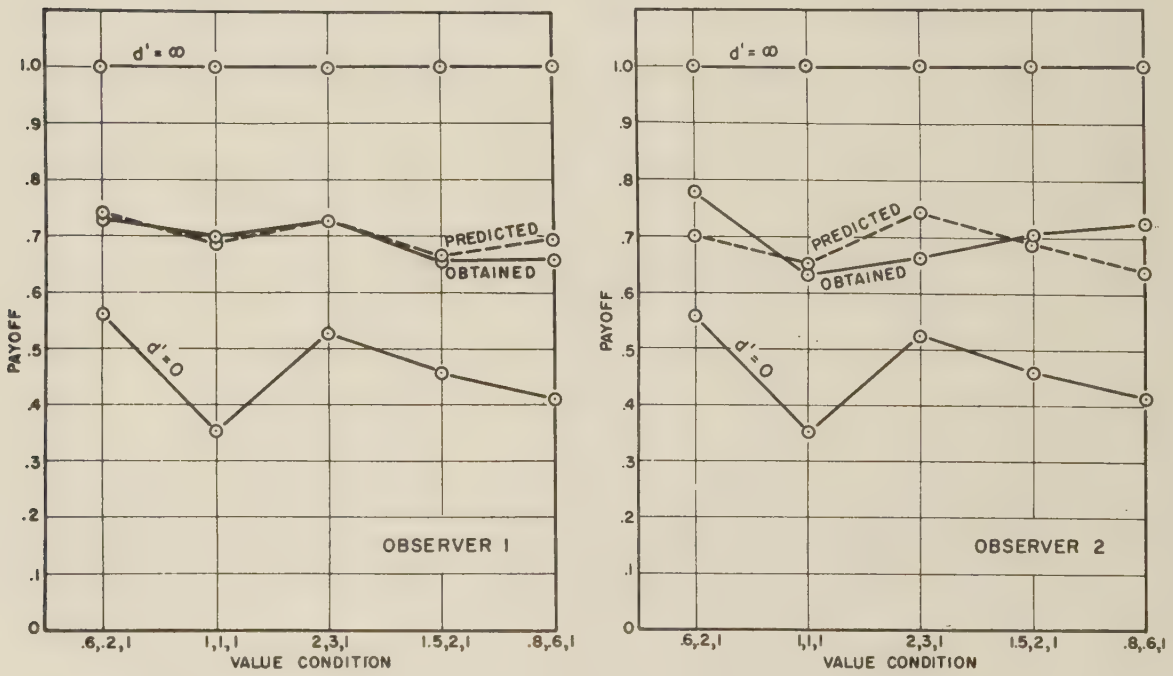


Fig. 9 - Various payoffs as a function of the value matrix.

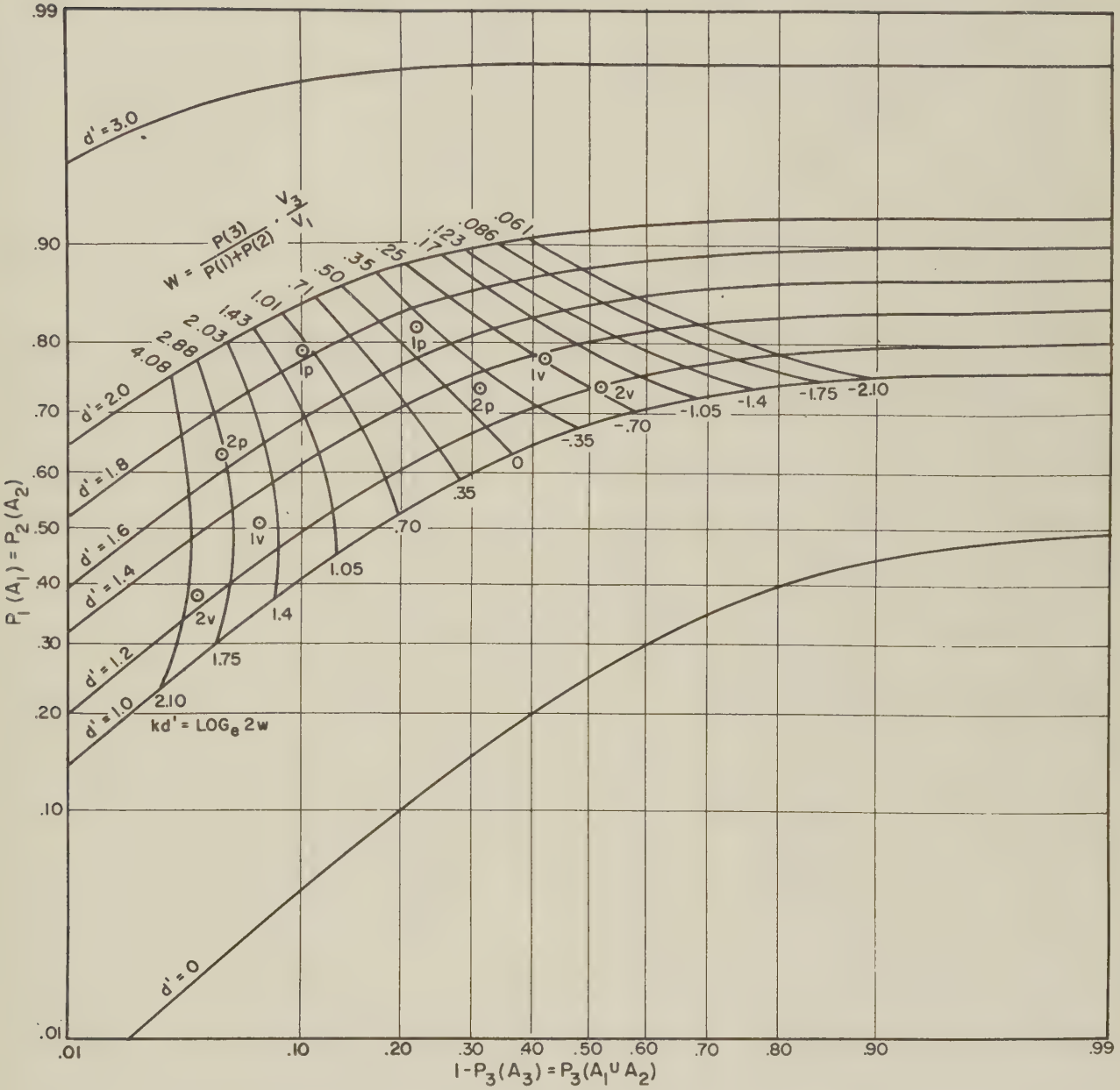


Fig. 10 - The data compared with the detectability curves specified by the decision-making theory.

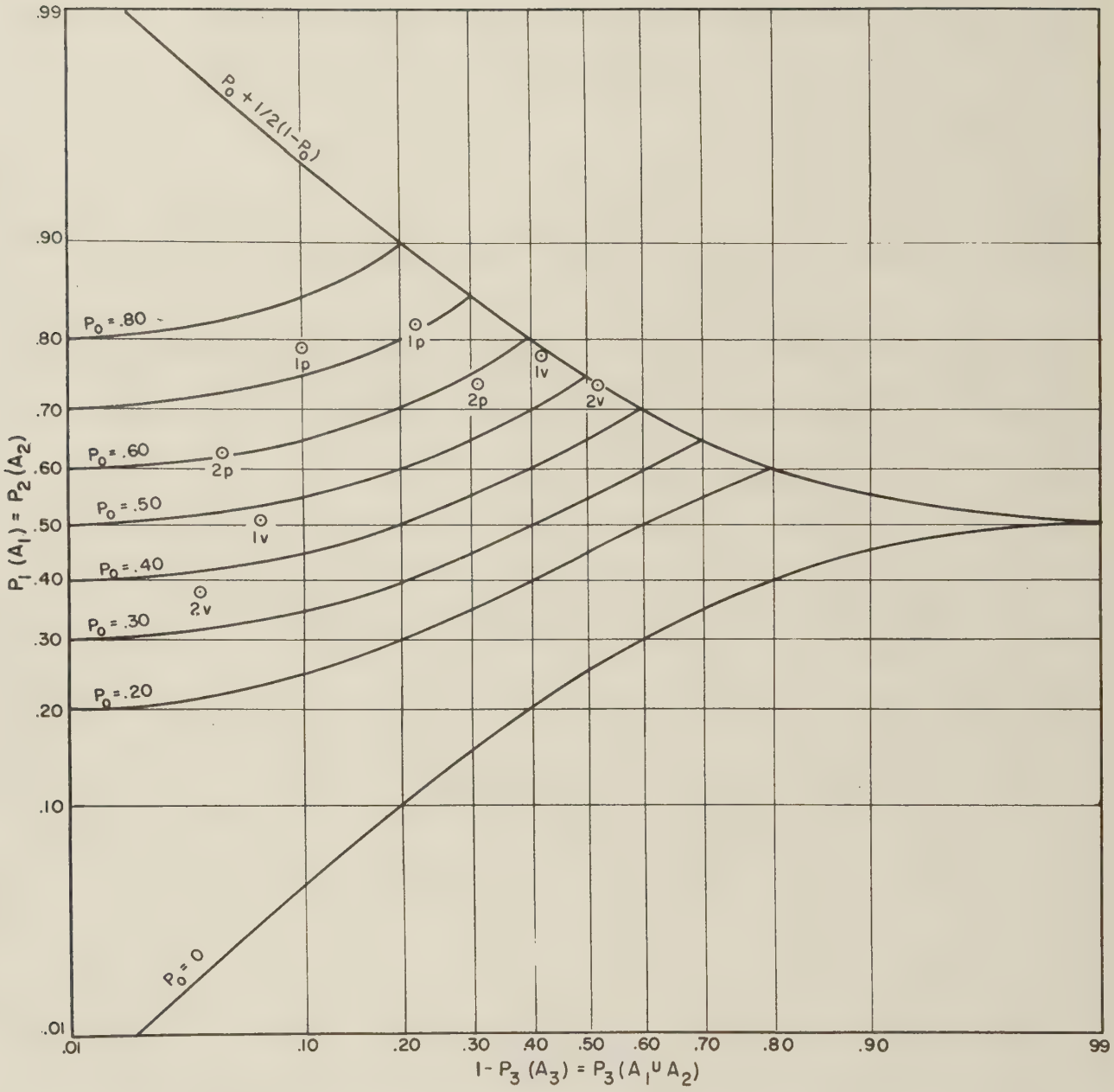


Fig. 11 - The data compared with the detectability curves specified by threshold theory.

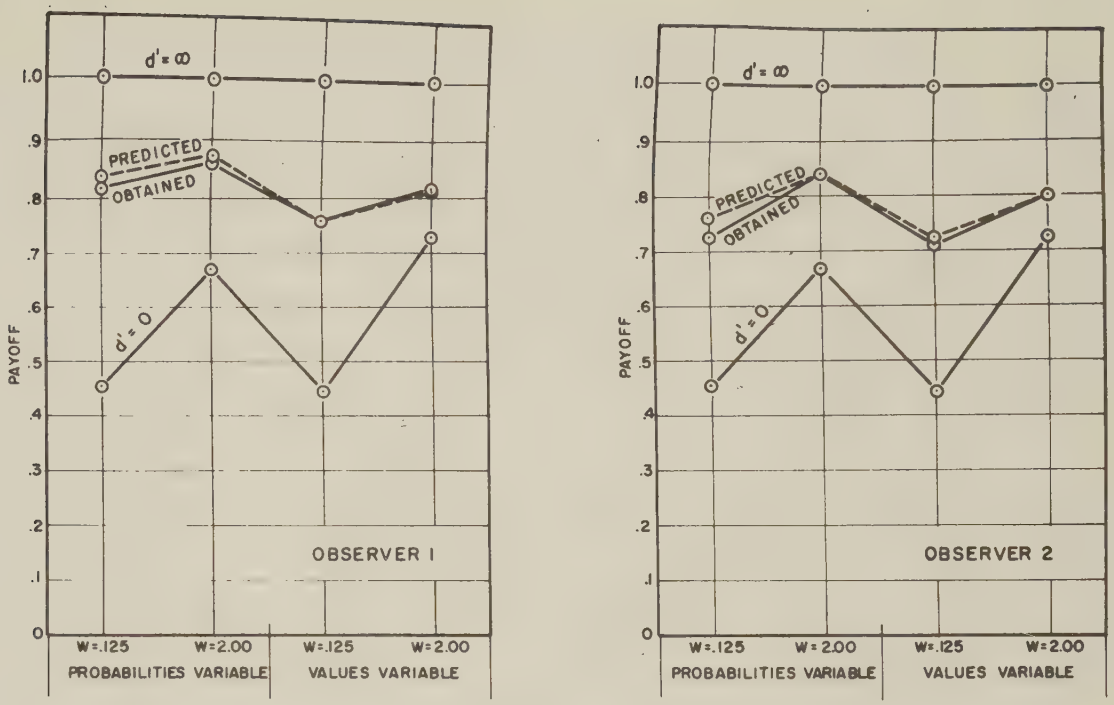


Fig. 12 - Various payoffs as a function of probabilities as values.

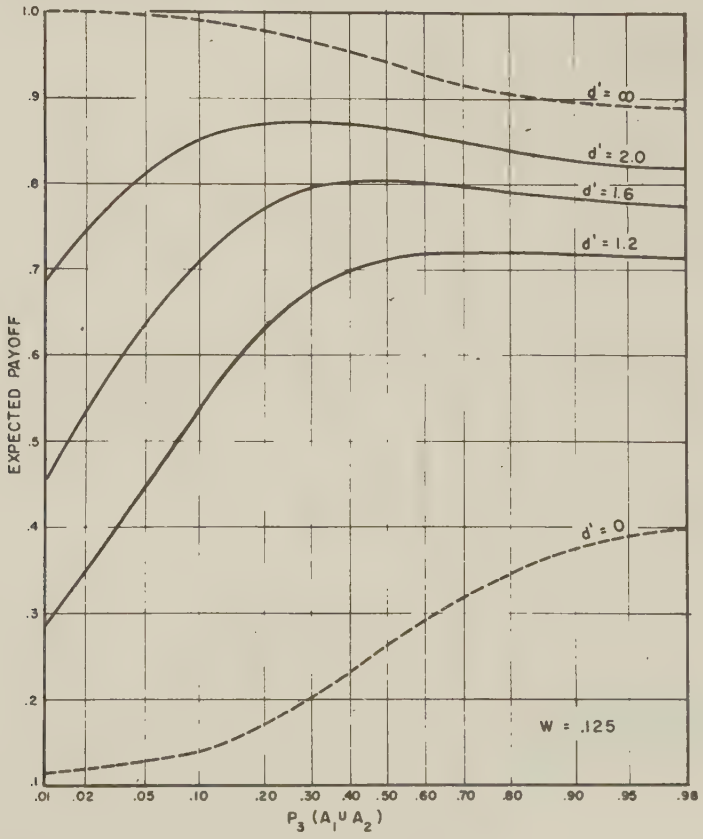


Fig. 13 - The expected payoff vs $P_3(A_1 \cup A_2)$ for w = .125

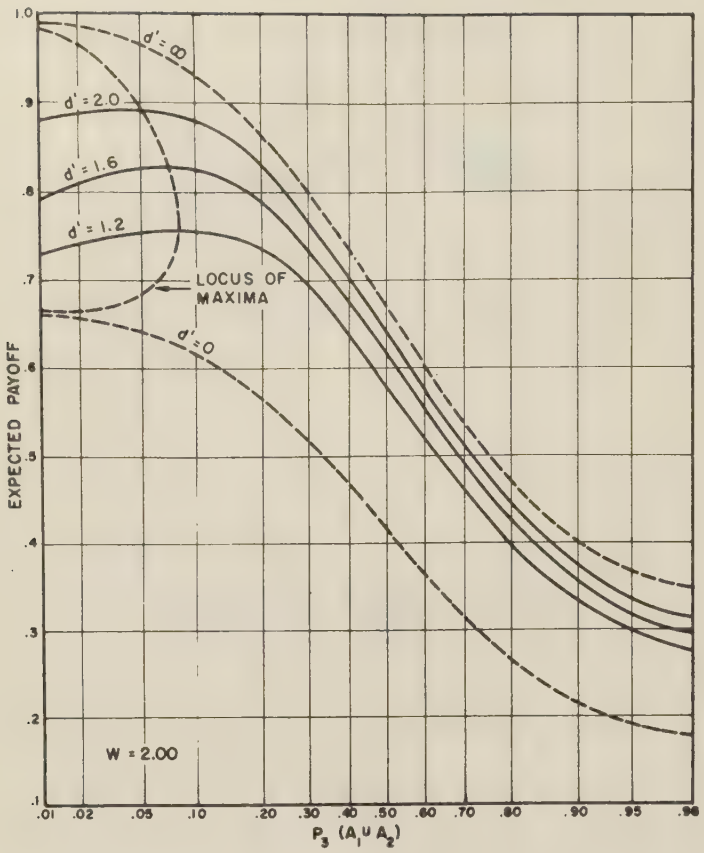


Fig. 14 - The expected payoff vs $P_3(A_1 \cup A_2)$ for w = 2.00.

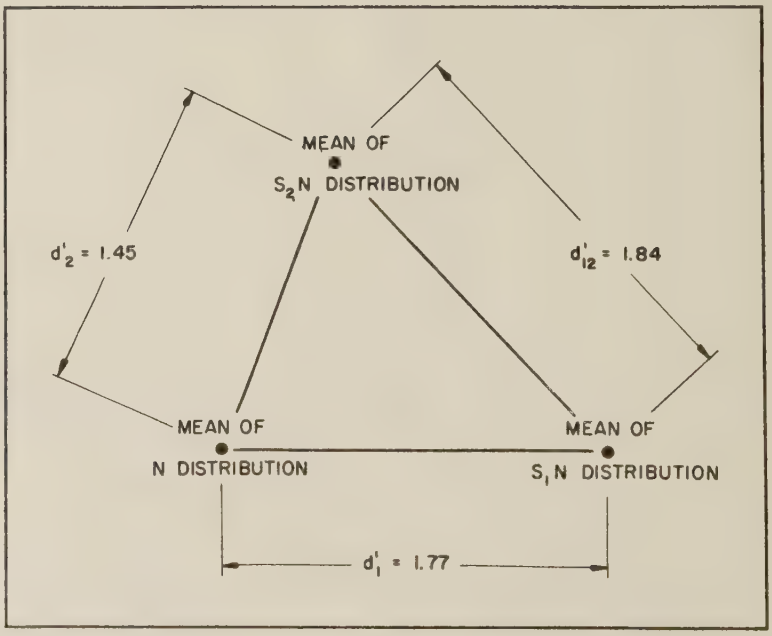


Fig. 15 - The geometrical model of the detection-and-recognition task.

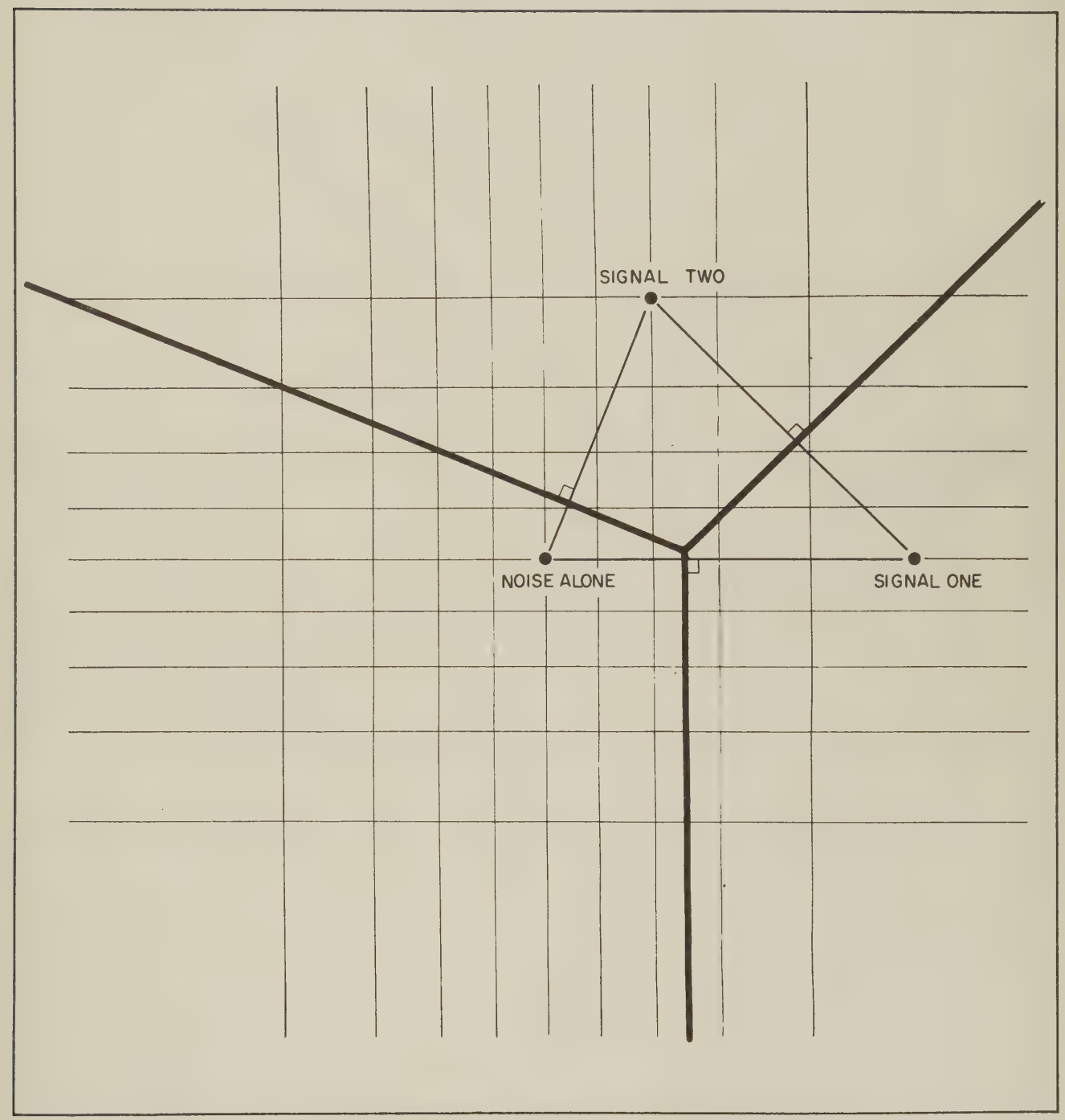


Fig. 16 - Noise alone presented.

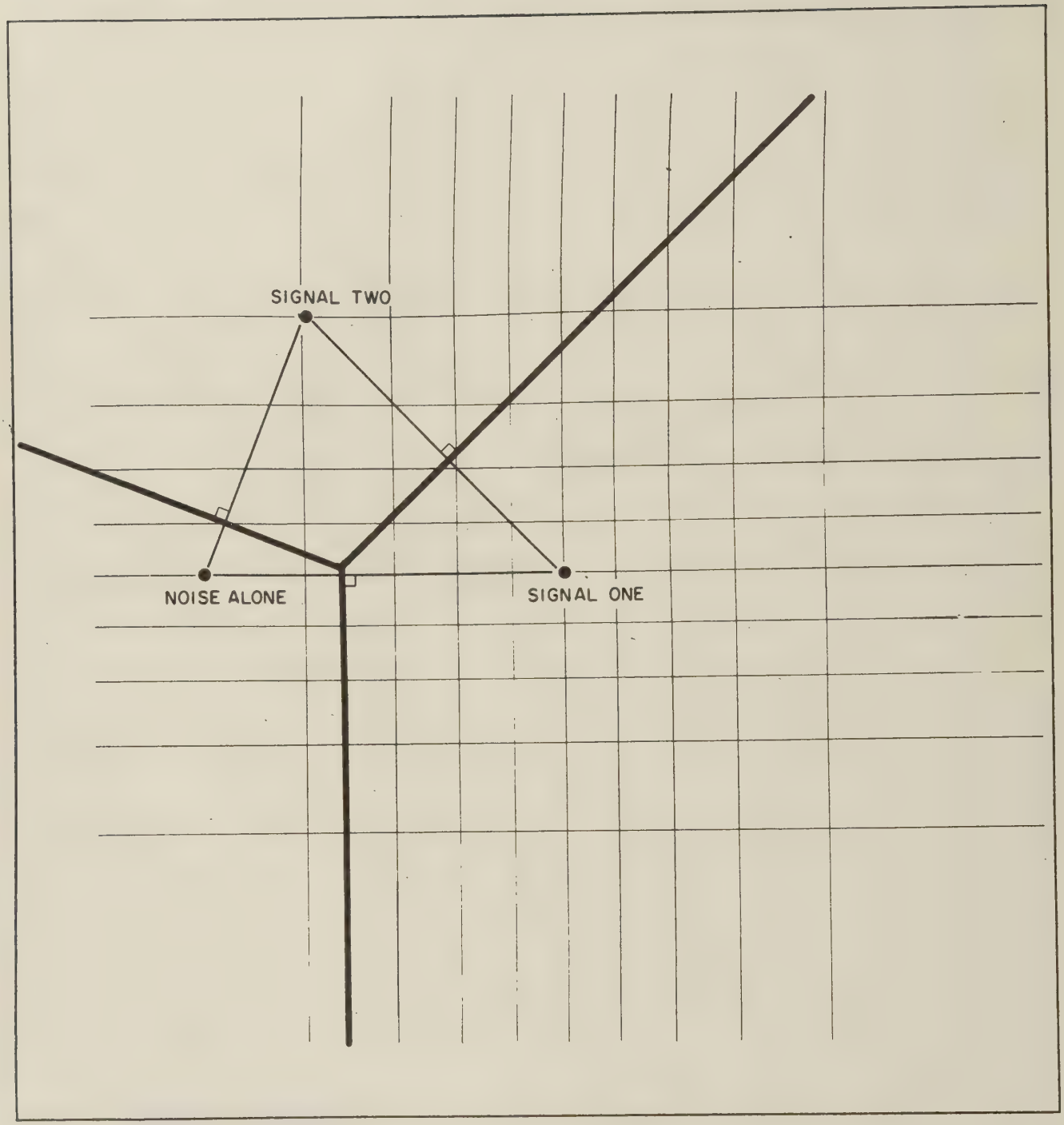


Fig. 17 - Signal one plus noise presented.

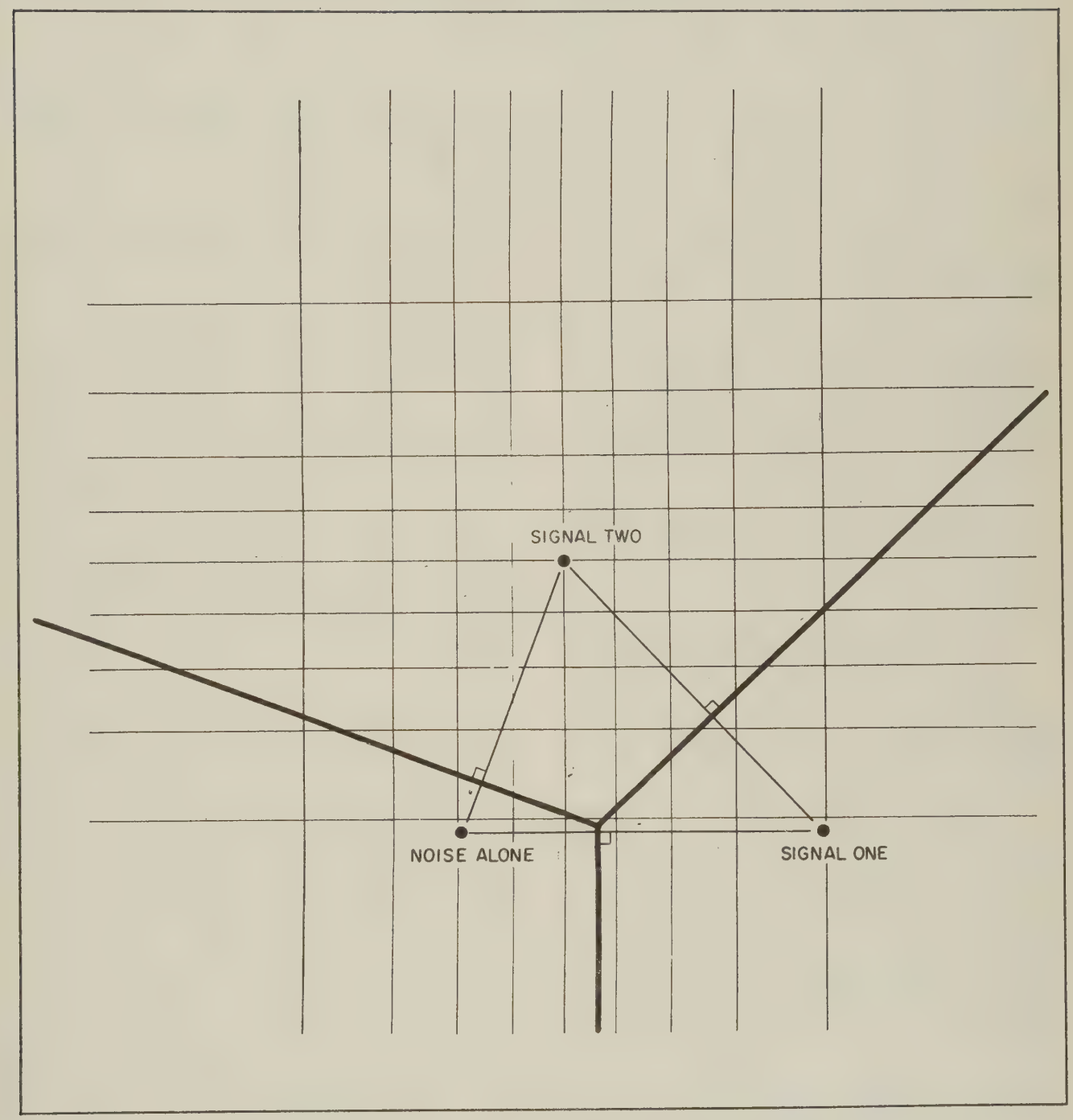


Fig. 18 - Signal two plus noise presented.

ON OPTIMUM NON-LINEAR EXTRACTION AND CODING FILTERS

A.V. Balakrishnan and R. Drenick*
GENERAL ENGINEERING DEVELOPMENT
RADIO CORPORATION OF AMERICA
CAMDEN, NEW JERSEY

SUMMARY

The problem of determining optimal non-linear least-square filters is solved for a class of stationary time series. This theory is then used as the basis for developing a band-width reduction scheme using non-linear encoding and decoding filters, for the same class of signals. A simple illustrative example is included.

This paper consists of two parts. In Part I we consider the problem of determining optimum non-linear (least-squares) extraction filters for a class of stationary time series. The results obtained in Part I provide the basis for a bandwidth-compression coding scheme which is discussed in Part II.

For the sake of simplicity, discrete parameter processes are considered.

I. OPTIMUM NON-LINEAR EXTRACTION FILTERS

Let $\{s_n\}$ be a strictly stationary** discrete parameter stochastic process identified as the "signal" and let $\{N_n\}$ be a second similar process, statistically independent of the first, identified as "noise". Let

$$x_n = s_n + N_n$$

Then $\{x_n\}$ is also a (strictly) stationary process. Let F be any translation-invariant operator on the part of $\{x_n\}$ so that

$$y_n = F[x_m,\ m \leq n] \tag{1}$$

The central problem with which we are concerned in this part of the paper is that of determining the form of the optimal F which minimizes the mean square error:

$$E[(s_n - y_n)^2]$$

where $E[\ ]$ denotes the expected value. [If, as is usual, ergodicity is assumed the phase averages may be replaced by time averages.]

Now it is well-known[1] that if the optimal estimate is denoted by s_n^+,

$$s_n^+ = E[s_n | x_m,\ m \leq n] \tag{2}$$

which thereby determines the optimal F as well. Further, the actual minimal mean square error itself is given by:

$$\text{minimal error} = E[(s_n)^2] - E[(s_n^+)^2] \ \ldots \tag{3}$$

The problem, however, is that of determining the optimal operator given by (2) in closed functional form. When F is restriced to be linear, complete solution is possible and if, in addition, ergodicity is assumed, can be obtained by working entirely with time averages, as Wiener does in his classic work.[2] With the linearity restriction removed, however, the complexity of the problem increases considerably, in common with most non-linear problems. Progress can nevertheless be made if, as a first step, the class of processes considered is suitably restricted, while still retaining engineering usefulness. Such a restriction we consider below.

It is well-known that any Gaussian process can be derived from white Gaussian noise [shot noise] by a linear filter. Extending this, we assume that the structure of the signal and noise processes to be considered in this paper, is such that they can both be derived from stationary "pure white" (but not necessarily Gaussian) primary processes by linear filters acting only on the past. We further assume, for the purposes of paper, that the linear filters differ only by a multiplicative constant. Under these conditions we have the representation:

$$\left.\begin{aligned} s_n &= \sum_0^\infty w_k\, \zeta_{n-k} \\ N_n &= \sum_0^\infty w_k\, \eta_{n-k} \end{aligned}\right\} \tag{4}$$

where we may further take without loss in generality

$$\sum_0^\infty w_k^2 = 1$$

Since $\{s_n\}$, $\{N_n\}$ are statistically independent, so are $\{\zeta_n\}$ and $\{\eta_n\}$, besides being pure white processes. A consequence of this representation is, as we shall see, that the optimal linear filter is reduced to a multiplicative constant.

* *Now with RCA, West Coast Engineering, Los Angeles.*

** *For the terminology used see reference 1.*

We note that if we let

$$z_n = \zeta_n + \eta_n ,$$

$\{z_n\}$ is also a pure white process, and further we have the representation

$$x_n = \sum_0^\infty w_K \; z_{n-K} \tag{5}$$

We next make the assumption that the transformation in (5) has an inverse. We have then the reciprocal relationship:

$$z_n = \sum_0^\infty g_K \; x_{n-K} \tag{6}$$

In order to obtain the functional form of the optimal operator, we begin by using the basic result (2)

$$s_n^+ = E[s_n | x_n, \; m \leq n]$$

To avoid index trouble, we now use the stationarity property. (Indeed, it may be noted that our representation (4) makes the processes ergodic.) Then it suffices to find*

$$s_o^+ = E[s_o | x_m, \; m \leq 0]$$

If we now use the fact that the transformation (5) is one-one, we have:

$$s_o^+ = E[s_o | z_m, \; m \leq 0]$$

the $\{z_m\}$ being determined in terms of the $\{x_m\}$ by relation (6). Again, since

$$s_o = \sum_0^\infty w_K \; \zeta_K ,$$

it follows that

$$E[s_o | z_m, \; m \leq 0] = \sum_0^\infty w_K \; E[\zeta_K | z_m, \; m \leq 0]$$

the convergence being in the mean and with probability one. Now

$$E[\zeta_K | z_m, \; m \leq 0] = E[\zeta_K | z_K]$$

Since ζ_K is independent of z_m for $m \neq K$.

Thus we have that:

$$s_o^+ = \sum_0^\infty w_K \; E[\zeta_K | z_K] \tag{7}$$

Again, stationarity permits us to write

$$E[\xi_K | z_K] = f(z_K) \tag{8}$$

So that we have finally,

$$s_o^+ = \sum_0^\infty w_K \; f(z_K) \tag{9}$$

and

$$s_n^+ = \sum_0^\infty w_K \; f(z_{n-K}) \tag{10}$$

* *The convergence properties involved here may be derived for instance from the Martingale theory of Doob, reference 1.*

At this point we can block-diagram the nonlinear filter represented by (10). Thus in Figure 1, the optimal filter consists of 3 sections, a zero-memory device sandwiched between two mutually reciprocal linear filters: The $\{x_n\}$ process is first equalized to have a flat spectrum, that is, to yield $\{z_n\}$. The zero memory device has transfer characteristic $f(\,.\,)$. The third filter is the reciprocal of the first and restores the spectrum to its original shape. It may be noted that filter memory is confined to the linear section.

Synthesis of The Zeromemory Filter

From (10) it is seen that the major problem in the synthesis of the optimal filter is the determination of the form of the function $f(\;)$. [It will be noted that determining $f(\;)$ is equivalent to solving the optimisation problem for pure white processes.] Actually the problem here is two-fold: One is to obtain $f(\;)$ in closed form and the second is to obtain a good approximation for it suitable for physical synthesis.

The first step in determining $f(.)$ is of course to specify the distributions of ζ_K and z_K. Here it has been found most convenient to assume that they can be approximated by taking a finite number of non-zero coefficients in a Gram-Charlier Expansion.[3] Of course, other representations are possible, but here again, our preference is guided by the needs in Part II. Before we use the Gram-Charlier representation, we note that

$$\begin{aligned} f(z_K) &= E[\zeta_K | z_K] \\ &= \int \zeta_K \; P(\zeta_K | z_K) \; d\zeta_K \\ &= \frac{\int s \; P_{\zeta_K}(s) \; P_{\eta_K}(z_{K-s}) d\zeta}{P_{z_K}(z_K)} \end{aligned} \tag{11}$$

To simplify (11) further, it is somewhat easier to work with characteristic functions. Let

$$C_\zeta(t) = \int exp \; it\zeta \; P_{\zeta_K}(\zeta) \; d\zeta$$

$$C_\eta(t) = \int exp \; it\eta \; P_{\eta_K}(\eta) \; d\eta$$

Then it follows that

$$f(z_K) = \frac{\int (exp - it \; z_K)(-i)\left[\frac{\partial}{\partial t} \; C_\zeta(t)\right] C_\eta(t) \; dt}{\int (exp - it \; z_K) \; C_\zeta(t) \; C_\eta(t) \; dt} \tag{12}$$

using independence of the ζ_K and η_K process and simple properties of convolutions.

The use of characteristic functions makes it particularly easy to work with Gram-Charlier expansions. For simplicity we assume zero means and let

$$\left.\begin{aligned} C_\zeta(t) &= \left[1 + \sum_{n=0}^{N} \frac{\alpha_n}{n!} (it)^n\right] \exp - \sigma_s^2 \frac{t^2}{2} \\ C_\eta(t) &= \left[1 + \sum_{n=3}^{N} \frac{\beta_n}{n!} (it)^n\right] \exp - \sigma_N^2 \frac{t^2}{2} \end{aligned}\right\} \quad (13)$$

where σ_s, σ_N are the variances of ζ_K and η_K (or of ζ_K and N_K) respectively. Substituting (13) into (12) we have:

$$f(z_K) = \frac{\displaystyle\int (-i)\left[-\sigma_s^2 t - \sigma_s^2 t \sum_3^N \frac{\alpha_n}{n!} (it)^n + i \sum_3^N \frac{\alpha_n}{(n-1)!} (it)^{n-1}\right] \left[1 + \sum_3^n \frac{\beta_n}{n!} (it)^n\right] \exp - \left(\frac{\sigma_\zeta^2 + \sigma_N^2}{2}\right) t^2 \exp - iz_K t \, dt}{\displaystyle\int \left[1 + \sum_{n=0}^{N} \frac{(\alpha_n + \beta_n)}{n!} (it)^n + \sum_{K=3}^{N} \sum_{l=3}^{N} \frac{\alpha_K \beta_l}{K!\, l!} (it)^{K+l}\right] \exp - \frac{(\sigma_s^2 + \sigma_N^2)}{\alpha} t^2 \exp - iz_K t \, dt} \quad (14)$$

Relation (14) can be further simplified in two ways, both useful. In the first, and obvious way, we let

$$\int (-it)^n \exp - iz\, t \exp - \frac{\sigma_s^2 + \sigma_N^2}{2} t^2 \, dt \; Exp \frac{z^2}{2(\sigma_s^2 + \sigma_N^2)} = \frac{\tilde{H}_n(z)}{(-1)^n} = \left[\exp + \frac{z^2}{2(\sigma_s^2 + \sigma_N^2)}\right] \frac{d}{dz^n} \exp - \frac{z^2}{2(\sigma_s^2 + \sigma_N^2)}$$

where $\tilde{H}_n(z)$ is the Hermite polynomial of degree n associated with the Gaussian of variance $(\sigma_s^2 + \sigma_N^2)$. Then $f(z)$ can be expressed as a ratio of these Hermite polynomials as

$$f(z) = \frac{\displaystyle\sum_{n=1}^{2N+1} \frac{a_n}{n!} \tilde{H}_n(z)}{1 + \displaystyle\sum_{n=3}^{2N} \frac{\gamma_n}{n!} \tilde{H}_n(z)} \quad (15)$$

where the coefficients involved are obtained by inspection from (14). In particular, we note that the coefficients γ_n in the denominator are the Gram-Charlier coefficients in the expansion of the distribution of z_K.

Unfortunately the denominator in (15) makes mechanization difficult and further is not sufficiently suggestive. For example, it does not provide any answer to a question such as: What is the best nth order approximation to the filter? For this purpose, we note first from (11) that $f(z)$ is of the form

$$f(z) = \frac{g(z)}{P(z)}$$

where the denominator is the distribution of z_K. Let $\{P_n(z)\}$ be the sequence of polynomials orthonormal[3] with respect to $P(z)$, so that

$$\int P_n(z)\, P_m(z)\, P(z)\, dz = \delta_n^m$$

Then $f(z)$ can be expanded in terms of these polynomials so that

$$f(z) = \sum_0^\infty c_n P_n(z) \ldots \quad (16)$$

where

$$c_n = \int P_n(z)\, g(z)\, dz \ldots \quad (17)$$

The convergence of (16) is both in the mean and with probability one, and further, $\sum_0^n c_n P_n(z)$ is the best approximation to $f(z)$ in the mean by an nth order polynomial.

Moreover, substituting (16) into (10), we obtain

$$s_n^+ = \sum_0^\infty w_K \left[\sum c_l P_l(z_{n-K})\right] \quad (18)$$

$$= \sum_0^\infty c_l \left[\sum w_K P_l(z_{n-K})\right] \quad (19)$$

where in the second form the lth term in the series yields the best lth order approximation to the filter. (It is interesting to note parenthetically that the filter given by (10) is already of class N_∞, in general, in the classification scheme of Zadeh,[4] while (19) explicitly provides a "power series" expansion for it.)

The minimal error given by (3) can again be expressed using (18) as:

$$\epsilon^2 = E[s^2] - E[(s^+)^2]$$

$$= \left[\sum_0^\infty w_K^2\right] \left[\sigma_s^2 - \sum_0^\infty c_l^2\right]$$

$$= \sigma_s^2 - \sum_0^\infty c_l^2 \ldots \quad (20)$$

Relation (20) makes clear how the error decreases as higher-order approximations are used.

The orthonormal polynomials $P_n(z)$ can be explicitly obtained in terms of the moments of $P(z)$ as given for instance by Cramer, reference 3, p. 132. In computing the coefficients, it is convenient to note that:

$$\int z^n g(z) dz = (-i)^n \frac{d}{dt^n}\left[\sum \frac{a_n}{n!} (it)^n \exp \frac{-t^2(\sigma_s^2 + \sigma_N^2)}{2}\right]\Bigg|_{t=0} \quad (21)$$

In particular, the first two polynomials can be easily seen to be:

$$P_0(z) = 1$$

$$P_1(z) = \frac{z}{\sqrt{\sigma_s^2 + \sigma_N^2}}$$

and

$$c_0 = 0$$

$$c_1 = \frac{\sigma_s}{\sqrt{\sigma_s^2 + \sigma_N^2}}$$

so that the best linear filter is given by

$$\frac{\sigma_s^2}{\sigma_s^2 + \sigma_N^2} \left[\sum_0^\infty w_K \, z_{n-K} \right] = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_N^2} \, x_n$$

with corresponding error

$$= \sigma_s^2 - \frac{\sigma_s^4}{\sigma_s^2 + \sigma_N^2}$$

$$= \frac{\sigma_s^2 \, \sigma_N^2}{\sigma_s^2 + \sigma_N^2}$$

The best linear filter is thus reduced a multiplicative constant. The computation of higher order terms in (18) in this generality is as laborious as unnecessary and the results for a specific example can be found in Part II.

2. OPTIMUM CODING AND DECODING FILTERS

The optimal non-linear filter theory outlined in Part I will now be used to develop a bandwidth-reduction coding scheme for the same class of signals, which is relatively simple to implement:

To be specific, we consider two statistically independent signals with identical statistics of all orders (such as, for instance, two independent long samples drawn from the same population), and show how these may be transmitted so as to occupy only as much frequency-band as required by each signal individually, and received subject to any specified degree of fidelity. The measure of fidelity adopted here in line with Part I is the mean square error.

A block-schematic of the system is given in Figure 2. Each signal is coded by an encoding filter at the transmitter prior to transmission over the same channel. The transmitted signal is thus the algebraic sum of the coded signals. At the receiver the operations of extraction of coded signals from the mixture and decoding to obtain the original signals are performed. The detailed structure of the various blocks is discussed below, but briefly, the coders serve to impart prescribed statistical characteristics to the signals so as to make the extraction as good as possible, while the decoders restore the original statistics of the individual signals, within the specified fidelity limits.

Let $\{s_n^1\}$, $\{s_n^2\}$ be the two signals, the superscripts identifying the signal, both of which are assumed to be representable as in (4) by

$$\left.\begin{aligned} s_n^1 &= \sum_0^\infty w_K \, \zeta_{n-K}^1 \\ s_n^2 &= \sum_0^\infty w_K \, \zeta_{n-K}^2 \end{aligned}\right\} \qquad (22)$$

where

$$\sum_0^\infty w_K^2 = 1$$

and

$$E[\zeta_n^1 \, \zeta_m^1] = E[\zeta_n^2 \, \zeta_m^2] = \delta_m^n$$

The two signals are assumed to have equal power and unit power is assumed for convenience. Identical statistics are obtained by taking all moments of ζ_n^1 and ζ_n^2 equal. If these signals were transmitted along the same channel without coding, the optimum least-squares filter would be linear, in fact would reduce to a multiplicative constant equal to one-half, leading to a mean square error of one-half.

The purpose of coding is to impart such characteristics to the signals as to improve the fidelity of reception. From our discussion in Part I we see that this may be done by non-linear filters which change the characteristics of the primary processes. Accordingly, we first equalize the signals $\{s_n^1\}$ and $\{s_n^2\}$ to obtain the $\{\zeta_n^1\}$ and $\{\zeta_n^2\}$ process, which it is possible to do for the class of signals we are considering. A simple zero-memory non-linear "encoding" device is used to charge the first-order moments of the equalizer signals in a way to be described presently. Finally the original spectrum is restored before transmission. Thus the coding filter illustrated in Figure 3 consists of 3 sections: an equalizer, a zero-memory device and a linear filter that restores the (spectral) shape of the spectrum. The problem then is of determining the characteristics of the zero-memory device.

Let us for the moment consider one of the signals, say $\{s_n^1\}$. Let the density of the equalized or primary process ζ_n^1 be expanded into a Gram-Charlier series:

$$P_{\zeta_n^1}(z) = \left[1 + \sum_3^N \frac{\alpha_n^1}{n!} H_n(z) \right] G(z) \qquad (23)$$

where

$G(z)$ is a Gaussian with unit variance and

and

$H_n(z)$ is the Hermite polynomials of order n associated with $G(z)$.

Since no change in power is desirable, as a result of coding, we can represent the density of the transformed variable ζ_n,

$$\zeta_n = g_1(\zeta_n^1)$$

with corresponding inverse transformation,

$$\zeta_n^1 = h(\zeta_n)$$

as

$$P_{\zeta_n}(z) = \left[1 + \sum_3^{\infty} \frac{\alpha_n}{n!} H_n(z)\right] G(z) \qquad (24)$$

where the α_n are still to be specified. But assuming for the moment they are the non-linear coder characteristic can be determined with the aid of (23) and (24) by the differential equation

$$G(y)\left[1 + \sum_3^{\infty} \frac{\alpha_n}{n!} H_n(y)\right]\frac{dy}{dx} = G(x)\left[1 + \sum_3^{\infty} \frac{\alpha_n^1}{n!} H_n(x)\right] \qquad (25)$$

where

$$y = g_1(x)$$

Solving this is routine and to avoid notational complexity we shall do this subsequently for a specific example:

The second signal is coded in a similar way, letting

$$\eta_n = g_2(\zeta_n^2)$$

so as to have

$$P_{\eta_n}(z) = \left[1 + \sum_3^{\infty} \frac{\beta_n}{n!} H_n(z)\right] G(z) \qquad (26)$$

The β_n's here are yet to be specified.

The design philosophy for determining the optimal α_n's and β_n's can take several forms. For the purposes of this paper, we have used the following: The α_n's and β_n's are so chosen as to make the mimimal error after extraction by the optimal filter (determined in Part I) as close to zero as possible. Since this means that the coded signals are recovered with arbitrarily small error by the extractor, the decoders are determined merely as the operational inverse of the coders. These considerations thus yield an overall synthesis method for the coding scheme.

A Simple Example—Gaussian Signals

We now abandon the level of generality of the preceding discussion and illustrate our ideas with a specific, though simple, example. The purpose is not so much to exhibit a finished system but rather to give a picture of what the operations look like in a simple case.

We assume that the signals $\{s_n^1\}$ and $\{s_n^2\}$ are Gaussian to begin with, so that the representations (4) and (6) are valid without any additional assumptions. We shall further confine ourselves to symmetric probability densities and coding operations which preserve symmetry of the probability densities throughout.

The simplest codes are obtained by considering the simplest departure from Gaussian. Thus, in the notation already used in this section, let

$$P_{\zeta_n}(z) = \left[1 + \frac{\alpha_4}{4!} H_4(z)\right] G(z) \qquad (27)$$

Here in order for the right-side to be non-negative, we must have

$$0 \leq \alpha_4 \leq 4$$

The corresponding differential equation for $g_1(.)$ becomes, with $y = g_1(x)$, and $x = h(x)$,

$$G(y)\left[1 + \frac{\alpha_4}{4!} H_4(y)\right]\frac{dy}{dx} = G(x)$$

which can be solved explicitly as:

$$x = h(y) = Erf^{-1}\left[Erf\ y + \frac{\alpha_4}{4!}(3y - y^3)\, G(y)\right] \qquad (28)$$

where

$$Erf\ y = \int_0^y G(t)dt$$

The resulting curve is plotted in Figure 4, using for α_4 the largest value possible, namely +4. The function $h(y)$ has an inflection point at $x = y = \sqrt{3}$; $x > y$ for $y < \sqrt{3}$, and $x < y$ for $y > \sqrt{3}$ and $x \to y$ as $y \to \infty$. x is an odd function of y and can be expanded in odd powers of y using (28).

Since the largest value of α_4 corresponds to the largest departure from Gaussian possible with the representation (27), we take $\alpha_4 = 4$ for our coder for signal #1. For signal #2, we must choose $\{\beta_n\}$ [or $g_2(.)$] so that the choice will lead to the smallest error in the extraction process. In this simple discussion, we bypass the variational problem involved here and suggest a heuristic argument. Thus, from the work in Part I, it would appear that to obtain a small error we have to impart a departure from Gaussian in the opposite sense, so to speak, to η_n, and choose $\beta_n \approx -\alpha_n$ [since for identical statistics $c_n = 0$ for $\eta_n \geq 2$ in (19)]. We do this by choosing for $g_2(.)$ the functional inverse of $g_1(.)$ so that

$$g_2(.) = h(.)$$

For this transformation, the resulting density of η_n can be readily deduced and is seen to be:

$$P_{\eta_n}(z) = \frac{G(z)}{1 + \frac{\alpha_4}{4!} H_4[g_1(z)]} \qquad (29)$$

To a first approximation $g_2(z) \sim z$, so that further

$$P_{\eta_n}(z) \sim \left[1 - \frac{\alpha_4}{4!} H_4(z)\right] G(z) \qquad (30)$$

which serves as an indication of the reasonableness of the choice. [An exact evaluation of the coefficients β_n is possible through the identity

$$\int \left[1 + \frac{\alpha_4}{4!} H_4(y)\right] G(y)\, Exp\ it\ f(y)\, dy = Exp - \frac{t^2}{2} \qquad (31)$$

from which it is also apparent that signal power is not preserved exactly, but only approximately so.] This choice of the 2nd coder is an advantage from a system point of view since the decoders do not involve a new design.

The probability distributions $P_{\zeta_n}(z)$ and $P_{\eta_n}(z)$ are plotted in Figures 5 and 6 respectively.

We next come to the extractor. The form of the extractor for the transmitted signal is determined either from (15) or (19). If we use (19) we can obtain $P_n(z)$ as in reference 3, noting now that all the odd moments vanish because of our symmetry assumption. We shall note here the expressions for the first three polynomials

$$P_0(z) = 1$$

$$P_1(z) = \frac{z}{\sqrt{2}}$$

$$P_3(z) = \frac{\mu_2 z^3 - \mu_4 z}{\sqrt{\mu_2(\mu_2 \mu_6 - \mu_4^2)}}$$

where μ_2, μ_4, μ_6 are the moments of $\zeta_n + \eta_n$ and can be evaluated in terms of the α_n's and β_n's. Thus

$$\mu_2 = 2$$

$$\mu_4 = 12 + \alpha_4 + \beta_4$$

$$\mu_6 = (\alpha_6 + \beta_6) + 30\mu_4 - 240$$

The corresponding coefficients are

$$c_0 = 0$$

$$c_1 = \frac{1}{\sqrt{2}}$$

$$c_3 = \frac{(\alpha_4 - \beta_4)}{\sqrt{(\mu_2\mu_6 - \mu_4^2)\mu_2}}$$

The extraction error using only the first 3 polynomials — or a "3rd order" coder — is thus

$$= 1 - \left[\frac{1}{2} + \frac{1}{2}\frac{(\alpha_4 - \beta_4)^2}{(2\mu_6 - \mu_4^2)}\right] \qquad (32)$$

Here $\alpha_4 = 4$; $\alpha_6 = 0$. β_4 and β_6 can be computed using (31) and making appropriate correction for unit power, but we can obtain a rough lower bound for the error by using the approximate form (30) and setting $\beta_6 = 0$, $\beta_4 = -4$ in (32). This yields the value 1/6.

This error is, however, too large and hence higher order coders and extractors have to be used to obtain a smaller error, if the assumed design philosophy is to be followed. We also note that the smaller the extraction error is, the closer the extractor-decoder scheme will be to the optimum, and the better the system performance.

The decoder being the inverse of the coder presents no additional design problem. The overall block diagram is shown in Figure 7, where both the spectrum shaping filter of the extractor and the equalizer of the decoder have been omitted since they are mutually reciprocal.

A similar extractor-decoder scheme can be used for the second signal. Here advantage may also be taken of the fact that the optimal estimate N_n^+ of any order is given by

$$N_n^+ = x_n - s_n^+$$

where s_n^+ is the optimal estimate of s_n to the same order, for possible simplification of the extractor.

REFERENCES

1. J.L. Doob, *Stochastic Processes*, John Wiley and Sons, New York, 1953.

2. N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley and Sons, New York, 1950.

3. H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

4. L.A. Zadeh, *A Contribution to the Theory of Non-Linear Systems*, Franklin Institute, 255. 387-408, 1953.

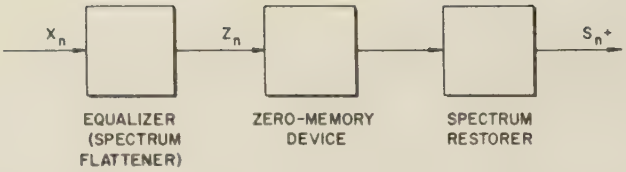


Fig. 1 - Optimal non-linear extraction filter.

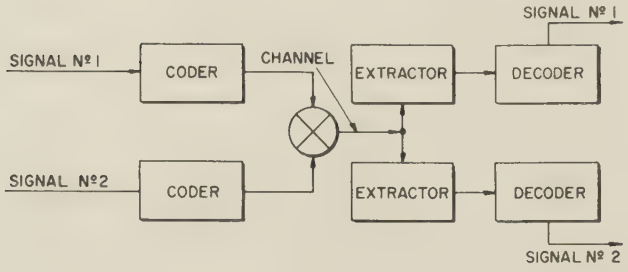


Fig. 2 - Block schematic of bandwidth-reduction system.

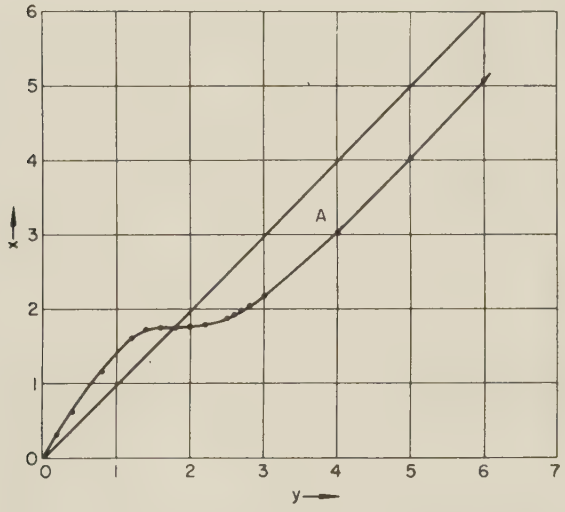


Fig. 4 - Zero-memory coder characteristics.

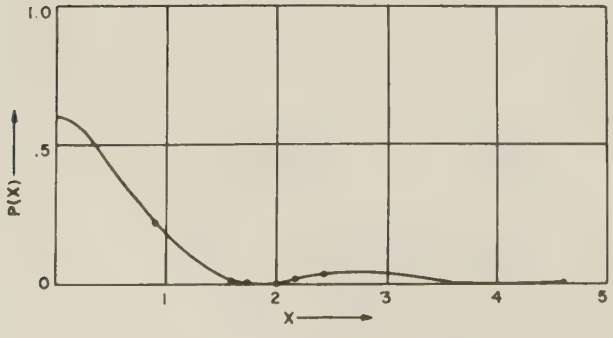


Fig. 5 - Probability density of ζ_{n1}

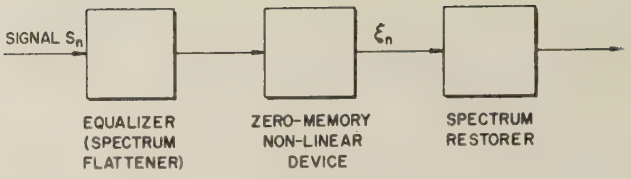


Fig. 3 - Structure of coding filters.

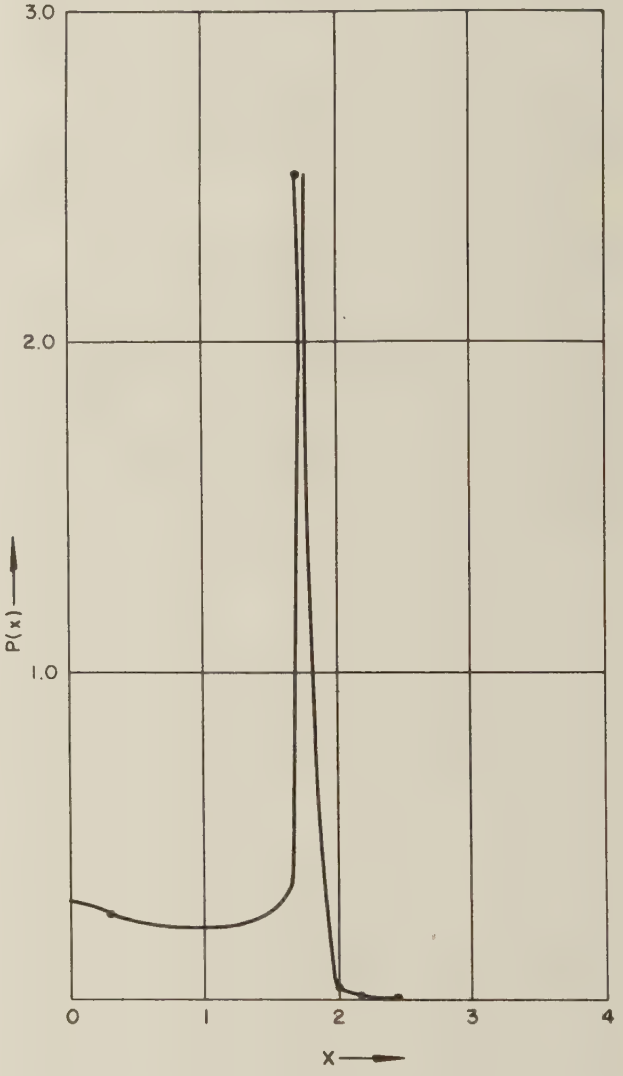


Fig. 6 - Probability density of η_n

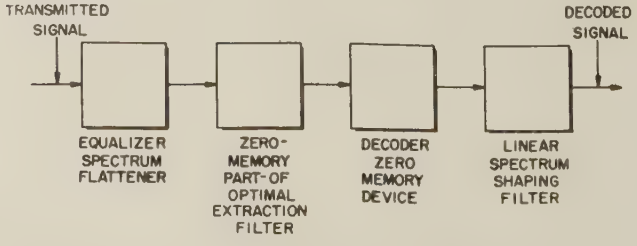


Fig. 7 - Block schematic of coder-extractor.

FINAL-VALUE SYSTEMS WITH GAUSSIAN INPUTS

by
Richard C. Booton, Jr.
Department of Electrical Engineering
and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract

A final-value system controls a response variable r(t) over a time interval (0,T) with the objective of minimizing the difference between a desired value ρ and the final response value r(T). An ensemble of situations is considered, and the system input i(t) and the desired response ρ are random variables that are statistically related. Physical limitations of the element being controlled result in a maximum-value constraint on the system velocity r'(t). Earlier results[1] suggest that a system consisting of an estimator followed by a "bang-bang" servo is approximately optimum. The estimator uses the input to produce an estimate ρ* of the desired response and the servo results in a system velocity as large in magnitude as possible and with the same sign as the difference ρ* - r. The present paper shows that this system is the true optimum when the joint distribution of the input and the desired response is Gaussian and the error criterion is minimization of the average of a nondecreasing function of the magnitude of the error.

Statement of the Problem

For mathematical convenience, the continuous time variable is replaced by a discrete set of n values, separated by an interval h where

$$h = \frac{T}{n} \tag{1}$$

The input is characterized by a sequence of values $i_1, i_2, \cdots, i_n$ and the response by a sequence $r_1, r_2, \cdots, r_n$. Each response r_k is a function only of $i_1, i_2, \cdots, i_k$. The constraint on velocity

$$|r'(t)| \leq V(t) \tag{2}$$

is replaced by

$$|r_k - r_{k-1}| \leq h\, V_k \tag{3}$$

The error criterion used is the minimization of a "cost" which is the average of a function of the final error. This cost is expressed in terms of the desired response ρ and the final response r_n as

$$C = E\left\{ f(\rho - r_n) \right\} \tag{4}$$

where E denotes the averaging operation (expectation). The function f is assumed to be an even uniminimal function, in the sense that

$$f(x) = f(-x) \tag{5}$$

and

$$f'(x) \geq 0 \qquad \text{for } x \geq 0 \tag{6}$$

The cost C can be expressed in terms of the joint probability density $p_{I\rho}$ of $i_1, i_2, \cdots, i_n$ and ρ as

$$C = \int \cdots \int f(\rho - r_n)\, p_{I\rho}(i_1, \cdots, i_n, \rho)\, di_1 \cdots di_n\, d\rho$$

The problem is to minimize the C given by this expression subject to the constraint (3).

Gaussian Variables

The density function $p_{I\rho}$ is Gaussian, and hence a set of independent variables can be introduced to simplify the calculations. With the conditional mean of ρ given $i_1, \cdots, i_k$ denoted by ρ_k^*, that is

$$\rho_k^* = E\left\{ \rho \,/\, i_1, \cdots, i_k \right\} \tag{7}$$

the variables defined by

$$\left.\begin{aligned} x_1 &= \rho_1^* \\ x_2 &= \rho_2^* - \rho_1^* \\ &\vdots \\ x_n &= \rho_n^* - \rho_{n-1}^* \\ x_{n+1} &= \rho - \rho_n^* \end{aligned}\right\} \tag{8}$$

are uncorrelated, as shown in Appendix A, and thus form a set of mutually independent Gaussian variables. With the probability density function of x_k denoted by p_k, the cost can be expressed as

$$C = \int \cdots \int f(x_{n+1} + \rho_n^* - r_n)\, p_{n+1}(x_{n+1}) \cdots p_1(x_1)\, dx_{n+1} \cdots dx_1 \tag{9}$$

The Optimum System

The cost expression can be rewritten as

$$C = \int \cdots \int f_n(\rho_n^* - r_n)\, p_n(x_n) \cdots p_1(x_1)\, dx_n \cdots dx_1 \tag{10}$$

*This work was supported in part by the Signal Corps, the Office of Scientific Research (Air Research and Development Command), and the Office of Naval Research, of the United States.

[1] R.C. Booton, Jr., "Optimum Design of Final-Value Control Systems", Proceedings of the Polytechnic Institute of Brooklyn Symposium on Nonlinear Network Analysis, 1956.

where

$$f_n(\rho_n^*-r_n) = \int f(x_{n+1}+\rho_n^*-r_n)p_{n+1}(x_{n+1})dx_{n+1} \quad (11)$$

Because f is an even uniminimal function and p_{n+1} is an even unimaximal function, the result in Appendix B implies that f_n is an even uniminimal function. The function $r_n(i_1, \cdots, i_n)$ minimizes C if for each set of values $(i_1, \cdots, i_n)$ the value of r_n is chosen to minimize f_n. Because the minimum value of $f_n(x)$ is achieved at x = 0 and f_n is nondecreasing away from this value, the optimum solution is to set $r_n = \rho_n^*$ if the constraint allows and otherwise to set $\rho_n^* - r_n$ as close to zero as allowed by the constraint. Thus, subject to the constraint

$$r_{n-1} - hV_n \leq r_n \leq r_{n-1} + hV_n \quad (12)$$

the value of r_n that minimizes f_n is

$$r_n = r_{n-1} + L_n(\rho_n^* - r_{n-1}) \quad (13)$$

where L_n is the limiter function defined by

$$\begin{aligned} L_n(x) &= -hV_n && \text{for } x < -hV_n \\ &= x && \text{for } |x| < hV_n \\ &= hV_n && \text{for } x > hV_n \end{aligned} \quad (14)$$

With this choice of r_n, f_n is a function of $\rho_n^* - r_{n-1}$. With this function denoted by g_n,

$$\begin{aligned} f_n(\rho_n^*-r_n) &= f_n\left[\rho_n^*-r_{n-1} - L_n(\rho_n^*-r_{n-1})\right] \\ &= g_n(\rho_n^* - r_{n-1}) \end{aligned} \quad (15)$$

This can be expressed as

$$f_n(\rho_n^*-r_n) = g_n(x_n+\rho_{n-1}^* - r_{n-1}) \quad (16)$$

Substitution of this relation into the cost expression (10) gives

$$C=\int\cdots\int g_n(x_n+\rho_{n-1}^*-r_{n-1})p_n(x_n)\cdots p_1(x_1)dx_n\cdots dx_1 \quad (17)$$

which can be written as

$$C=\int\cdots\int f_{n-1}(\rho_{n-1}^*-r_{n-1})p_{n-1}(x_{n-1})\cdots p_1(x_1) dx_{n-1}\cdots dx_1 \quad (18)$$

where

$$f_{n-1}(\rho_{n-1}^*-r_{n-1})=\int g_n(x_n+\rho_{n-1}^*-r_{n-1})p_n(x_n)dx_n \quad (19)$$

The function g_n is an even uniminimal function and hence, by the result of Appendix B, the function f_{n-1} also is an even uniminimal function.

The reasoning used to determine r_n as (13) can be repeated to determine a similar relation for r_{n-1}, and the entire procedure can be repeated so that, by induction, each response value r_k is given by an expression

$$r_k = r_{k-1} + L_k(\rho_k^* - r_{k-1}) \quad (20)$$

where

$$\begin{aligned} L_k(x) &= -hV_k && \text{for } x < -hV_k \\ &= x && \text{for } |x| < hV_k \\ &= hV_k && \text{for } x > hV_k \end{aligned} \quad (21)$$

The expression (20) can be written in the form

$$\frac{r_k - r_{k-1}}{h} = \frac{1}{h} L_k(\rho_k^* - r_{k-1}) \quad (22)$$

In the limit as h approaches zero, this becomes

$$\frac{d}{dt} r(t) = V(t) \operatorname{sgn}\left[\rho^*(t) - r(t)\right] \quad (23)$$

where

$$\left.\begin{aligned} \operatorname{sgn} x &= -1 && \text{if } x < 0 \\ &= 0 && \text{if } x = 0 \\ &= 1 && \text{if } x > 0 \end{aligned}\right\} \quad (24)$$

This is the principal result of the paper.

Appendix A

The purpose of this appendix is to demonstrate that the variables x_k defined by (8) are uncorrelated. Each conditional mean ρ_k^* is the best mean-square estimate of ρ given $i_1, \cdots, i_k$ and the general properties of mean-square estimation imply that

$$E\left\{(\rho - \rho_k^*)\lambda_k\right\} = 0 \quad (A\text{-}1)$$

where λ_k is any function of $i_1, \cdots, i_k$. In particular,

$$E\left\{(\rho - \rho_n^*)x_k\right\} = 0 \qquad \text{for } k = 1, \cdots, n \quad (A\text{-}2)$$

because each x_k is a function of the input values.

Use of

$$E\left\{(\rho - \rho_n^*)(\rho_n^* - \rho_{n-1}^*)\right\} = 0 \quad (A\text{-}3)$$

shows that

$$E\left\{(\rho-\rho_{n-1}^*)^2\right\} = E\left\{(\rho-\rho_n^*)^2\right\} + E\left\{(\rho_n^*-\rho_{n-1}^*)^2\right\} \quad (A\text{-}4)$$

This expression and the fact that ρ_{n-1}^* is the best estimate of ρ imply that ρ_{n-1}^* is the best estimate of ρ_n^* and

$$E\left\{(\rho_n^*-\rho_{n-1}^*)x_k\right\} = 0 \qquad \text{for } k=1,\cdots,n-1 \quad (A\text{-}5)$$

By induction, each ρ_k^*, for k = 1, ...,n-1, can be shown to be the best estimate of ρ_{k+1}^* and thus

$$E\left\{(\rho_{k+1}^*-\rho_k^*)x_k\right\} = 0 \text{ for } k=1,\cdots, n-1 \quad (A\text{-}6)$$

The equations (A-2) and (A-6) are equivalent to

$$E\left\{x_j x_k\right\} = 0 \quad \text{for all } j,\ k=1,\cdots,n+1 \qquad \text{(A-7)}$$

Appendix B

The purpose of this appendix is to show that the function H defined by

$$H(y) = \int_{-\infty}^{\infty} F(x+y)G(x)\,dx \qquad \text{(B-1)}$$

where

$$\left.\begin{aligned} &F(x) = F(-x) \\ &F'(x) \geq 0 \quad \text{for } x \geq 0 \end{aligned}\right\} \qquad \text{(B-2)}$$

and

$$\left.\begin{aligned} &G(x) = G(-x) \\ &G'(x) \leq 0 \quad \text{for } x \geq 0 \end{aligned}\right\} \qquad \text{(B-3)}$$

is even and

$$H'(y) \geq 0 \quad \text{for } y \geq 0 \qquad \text{(B-4)}$$

By a change of variable, (B-1) can be written as

$$H(y) = \int_{-\infty}^{\infty} F(-x+y)G(-x)\,dx \qquad \text{(B-5)}$$

Because F and G are even

$$H(y) = \int_{-\infty}^{\infty} F(x-y)G(x)\,dx \qquad \text{(B-6)}$$

and thus

$$H(y) = H(-y) \qquad \text{(B-7)}$$

Differentiation of (B-1) yields

$$H'(y) = \int_{-\infty}^{\infty} F'(x+y)G(x)\,dx \qquad \text{(B-8)}$$

By a change of variable,

$$H'(y) = \int_{-\infty}^{\infty} F'(x)G(x-y)\,dx \qquad \text{(B-9)}$$

which can be rewritten as

$$H'(y) = \int_{0}^{\infty} F'(x)G(x-y)\,dx + \int_{-\infty}^{0} F'(x)G(x-y)\,dx \qquad \text{(B-10)}$$

A change of variable in the second integral gives

$$H'(y) = \int_{0}^{\infty} F'(x)G(x-y)\,dx + \int_{0}^{\infty} F'(-x)G(-x-y)\,dx \qquad \text{(B-11)}$$

Because F' is odd and G is even

$$H'(y) = \int_{0}^{\infty} F'(x)\left[G(x-y)-G(x+y)\right]dx \qquad \text{(B-12)}$$

For $y \geq 0$ and $x \geq 0$

$$|x-y| \leq x+y \qquad \text{(B-13)}$$

and

$$G(x-y) \geq G(x+y) \qquad \text{(B-14)}$$

This inequality and (B-2) simply that the integral in (B-12) is non-negative and thus (B-4) holds.

AN EXTENSION OF THE MINIMUM MEAN SQUARE PREDICTION-THEORY FOR SAMPLED INPUT SIGNALS

Marvin Blum
CONVAIR
A Division of General Dynamics Corporation
San Diego, California

Abstract

A method is developed for finding the ordinates of a digital filter which will produce a general linear operator of the signal S(t) such that the mean square error of prediction will be a minimum. The input to the filter is sampled at intervals Δt. The samples contain stationary noise $N(j\Delta t)$, a stationary signal component, $M(j\Delta t)$, and a nonrandom signal component,

$$P(j\Delta t) = \sum_{k=o}^{n} a_k P_k (j\Delta t)$$

where the subset of nonrandom functions $P_k(t)$ are known a priori, but the parameter vector $a = (a_o, a_1, \ldots a_n)$ need not be.

The solution is obtained as a matrix equation which relates the ordinates of the digital filter to the autocorrelation properties of M(t) and N(t) and the nature of the prediction operation.

Introduction

The central problem in prediction lies in the fact that when a filter operates on an input signal to produce a desired output, there is usually random noise superimposed on the message which prevents determination of the desired output without error. It may be desirable therefore, to select a filter which is optimum in the following sense: The filter will produce an output whose ensemble average value is equal to the ensemble average value of the desired signal, and the mean square difference between the actual output of the filter and the desired output will be a minimum.

Norbert Wiener[1] considered this problem in his classical work on the subject. He considered the input function to be a continuous random stationary time series and the noise to be another continuous random stationary time series, and derived the solution for an optimum filter in the form of the Wiener-Hopf integral equation. It was required that the filter have a semi-infinite memory, e.g., the filter extended in the time domain from 0 to ∞. The work of Phillip and Weiss[2] extended the theory to include a nonrandom input signal in the form of a polynomial of known degree. A further extension was made by Zadeh and Ragazzini[3] in that they derived the equations for the optimum filter for a more complicated model. In their model the message signal consisted of a random stationary signal plus a polynomial of known degree, while the noise was random and stationary. In these three theories the autocorrelation of the noise and stationary signal are presumed to be known. In the last two the memory of the filter is finite, e.g., the impulsive admittance of the filter is zero outside some finite interval.

In a recent paper by Lees,[4] the solution to the Zadeh-Raggazini prediction model for sampled data was presented. The solution is obtained in the form of a weighing function which is piecewise analytical. The output of this filter is continuous as opposed to the output at sampled points only, of the digital filter.

However, by an extension of the prediction of the digital filter from a fixed parameter concept to a variable prediction, one can obtain the same results as shown by Lees. The extension of the digital filter to the piecewise analytic filter is presented for the more general input model considered in this paper.

A simplified solution for the piecewise analytic filter is presented where the input is a polynomial of degree n plus white noise, and the desired output is the predicted value of the K^{th} derivative of the input. These results are presented in the appendix.

In considering the solution of the prediction problem for discrete data, the similarity between the filter problem and the problems of curve fitting becomes apparent. In a paper by A. C. Aitken[5] the problem of the best linear transformation on a set of observations to fit a general class of nonrandom functions to the data is considered. His results are easily interpreted in

terms of a digital linear filter which smooths the input data. The importance of the approach is that it admits of a much broader class of nonrandom inputs. One can use an input (subject to some general restrictions stated within) which is a linear function of known functions without knowing the particular linear relationship.

In this paper the optimum digital linear filter is determined for the following conditions:

(a) The input signal to the digital filter consists of a stationary component plus a nonrandom component (from a more general class of nonstationary components than polynomials) plus random noise -- sampled at equidistant intervals.

(b) The prediction is a generalized linear operator of the message components. A particular application is made to derivative predictions.

The solution takes these forms:

(a) A matrix equation is derived which relates the ordinates of the weighting function to the stationary correlation properties of both signal and noise, the prediction operation, and the values of the n+1 subfunctions sampled at the m+1 equidistant points.

(b) The solution requires a weighting function with a finite memory, e. g., the discrete impulsive admittance of the filter is required to vanish outside an interval $0 \leqq t \leqq m\Delta t$ where Δt is the sampling interval and $m\Delta t$ is the finite memory of the filter.

(c) The smoothing operation on the data is interpreted as a dynamic curve-fitting procedure which yields minimum variance estimates as a series of outputs in real time. The first output becomes available when the first full series of m+1 data points are in the filter.

As succeeding data points become available, an optimum estimate can be obtained by a linear weighting of the most recent data point and the m previous data points. This sliding arc procedure requires in general a time-varying weighting sequence, but for a special subclass of the input function, the weighting sequence is time-invariant, e. g., the same coefficients in the weighting sequence multiply the input data at a fixed lag with respect to the most recent data point, even though the data changes with changing time.

Definition of Input Model and Weighting Sequence

Let the input to the filter be

$$e(t) = S(t) + N(t) \qquad (1$$

where

$$S(t) = P(t) + M(t) \qquad (2$$

and P(t) is the nonrandom part of the signal, M(t) is the stationary part of the signal, and N(t) is the stationary noise.

Let E(z) be defined as the expected value of z, or the ensemble average of z.

Then it is assumed that

$$E^{\dagger}\left[M(t)\right] = E\left[N(t)\right] = 0 \qquad (3$$

for all t, and that

$$P(t) = \sum_{k=0}^{n} a_k P_k(t). \qquad (4$$

That is, P(t) can be represented as a linear combination of the subset of n + 1 functions $P_k(t)$ where k = 0, 1, 2, . . n. Knowledge of the functions $P_k(t)$ is required a priori, but the value of the parameter vector $(a) = [a_0, a_1, a_2, \ldots a_n]$ need not be known. From equation 1 and 2 the sample input to the filter at time $t = (j\Delta t)$ is given by

$$e(j\Delta t) = N(j\Delta t) + M(j\Delta t) + P(j\Delta t). \qquad (5$$

Let W'(t) be the impulsive admittance of the filter. Then the output of the digital filter at time $t = (m+u)\Delta t$ where u = 0, 1, 2, . . .) is given by

$$e^*\left[(m+u)\Delta t\right] = \sum_{j=0}^{\infty} \Delta t W'(j\Delta t) e\left[(m+u-j)\Delta t\right] . \qquad (6$$

Let $W_j \equiv \Delta t W'(j\Delta t)$, then

$$e\left[(m+u-j)\Delta t\right] \equiv e_{m+u-j} = N_{m+u-j} + M_{m+u-j} + P(m+u-j),$$

and

$$e^*\,(m+u)\Delta t \equiv e^*_{m+u} .$$

Let us assume that m>n and that $W_j = 0$ wherever j<0 or j>m. Then equation 6 can be written as

† In this case E [M(†)] refers to the ensemble average of M at time †.

$$e^*_{m+u} = \sum_{j=0}^{m} W_j \left[N_{m+u-j} + M_{m+u-j} + P(m+u-j) \right] . \quad (7$$

The vector $(W_u) = \left[W_{0,u}, W_{1,u}, W_{2,u}, \cdots W_{m,u} \right]$ will be noted as the weighting sequence.

In the subsequent solution for the weighting sequence the parameter u will be taken equal to zero and deleted from the notation. Where necessary certain matrixes will be defined for u not equal to zero. The solution for u equal to zero will be presented and then the modification of the solution for u not equal to zero will be given.

Generalized Desired Output

Let $S^*(m\Delta t)$ be the desired output. Since the desired output is defined as a fixed linear operation of the input one may relate the impulsive response of an ideal predictor $k'(t)$, to the input and output relationships, by an equation of the form

$$S^*(t) = \int_{-\infty}^{+\infty} k'(\tau) S(t-\tau) d\tau . \quad (8$$

For $t = m\Delta t$, using equations 2 and 8, one obtains

$$S^*(m\Delta t) = \int_{-\infty}^{+\infty} k'(\tau) P(m\Delta t-\tau) d\tau + \int_{-\infty}^{+\infty} k'(\tau) M(m\Delta t-\tau) d\tau . \quad (9$$

Using equation 4 one obtains

$$S^*(m\Delta t) = \sum_{k=0}^{n} a_k \int_{-\infty}^{+\infty} k'(\tau) P_k(m\Delta t-\tau) d\tau + \int_{-\infty}^{+\infty} k'(\tau) M(m\Delta t-\tau) d\tau . \quad (10$$

Let

$$Q_k = \int_{-\infty}^{+\infty} k'(\tau) P_k(m\Delta t-\tau) d\tau \quad (11$$

and

$$Q_{k,u} = \int_{-\infty}^{+\infty} k'(\tau) P_k \left[(m+u)\Delta t-\tau \right] d\tau . \quad (12$$

The matrixes $|Q|$ and $|Q_u|$ are defined as

$$|Q| = \begin{vmatrix} Q_0 \\ Q_1 \\ \vdots \\ Q_n \end{vmatrix} \qquad |Q_u| = \begin{vmatrix} Q_{0,u} \\ Q_{1,u} \\ \vdots \\ Q_{n,u} \end{vmatrix} . \quad (13$$

Let

$$M^0 \equiv \int_{-\infty}^{+\infty} k'(\tau) M(m\Delta t-\tau) d\tau . \quad (14$$

Then

$$S^*(m\Delta t) = (a)\,|Q| + M^0 \quad (15$$

where $(a)^\dagger$ is the row matrix $(a) = (a_0, a_1, a_2, \ldots a_n)$.

As an example, consider the relationship

$$S^*(m\Delta t) = \frac{d}{dt} s(t) \Big|_{t = (m+\alpha)\Delta t} . \quad (16$$

Equation 16 reads, the desired output in real time at $t = m\Delta t$ is the value of the first derivative of the input function evaluated at $t = (m+\alpha)\Delta t$. Equation 11 may be written for equation 16 as

$$Q_k = \frac{d}{dt} P_k(t) \Big|_{t = (m+\alpha)\Delta t} \quad (17$$

and equation 14 becomes

$$M^0 = \frac{d}{dt} M(t) \Big|_{t = (m+\alpha)\Delta t} . \quad (18$$

Equations of Constraint on (W)

Let the error be defined as

$$\epsilon_m = e^*(m\Delta t) - S^*(m\Delta t) \quad (19$$

where ϵ_m is the difference between the actual output $e^*(m\Delta t)$ and the desired output $S^*(m\Delta t)$ at time $t = m\Delta t$. One seeks the optimum weighting function (W) which is inferred from the following conditions on ϵ_m:

† ( ) indicates a row matrix, while || indicates a column matrix.

$$E(\epsilon_m) = 0 \qquad (20\ (a$$

$$E(\epsilon_m^2) = \text{minimum.} \qquad (b$$

Equation 19 and requirement 20a infer that

$$E\left[e^*(m\Delta t)\right] = E\left[S^*(m\Delta t)\right] . \qquad (21$$

Using equation 15 one finds

$$E(S^*m\Delta t) = (a)\ |Q| . \qquad (22$$

Using equation 3, 7, and 21 leads to

$$E(e^*(m\Delta T)) = \sum_{j=0}^{m} W_j\, P(m-j) . \qquad (23$$

Substituting equation 4 into equation 23 one obtains

$$E\left[e^*(m\Delta t)\right] = \sum_{j=0}^{m}\sum_{k=0}^{n} W_j\, P_k(m-j)\, a_k . \qquad (24$$

Equation 24 can be written in matrix form as follows:

$$E\left[e^*(m\Delta t)\right] = (a)\ P\ |W| \qquad (25$$

where

$$(a) = (a_0,\ a_1,\ \ldots\ a_n) ,$$

and

$$|W| = (W)',\ \text{e.g., transpose of}\ W = \begin{vmatrix} W_0 \\ W_1 \\ \cdot\cdot \\ \cdot\cdot \\ W_m \end{vmatrix} ,$$

and

$$P = \begin{vmatrix} P_0(m) & P_0(m-1) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_0(0) \\ P_1(m) & P_1(m-1) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_1(0) \\ \cdot\cdot & \cdot\cdot & & \cdot\cdot & \cdot\cdot & \\ \cdot\cdot & \cdot\cdot & & \cdot\cdot & \cdot\cdot & \\ P_n(m) & P_n(m-1) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_n(0) \end{vmatrix}$$

and

$$P_u = \begin{vmatrix} P_0(m+u) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_0(u) \\ P_1(m+u) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_1(u) \\ \cdot\cdot & & \cdot\cdot & \cdot\cdot & \\ \cdot\cdot & & \cdot\cdot & \cdot\cdot & \\ P_n(m+u) & \cdot\cdot & \cdot\cdot & \cdot\cdot & P_n(u) \end{vmatrix} .$$

Note: A prime indicates that the transpose of a matrix has been taken.

From equations 21, 22, and 25 it follows that,

$$(a)\ |Q| = (a)\ P\ |W| . \qquad (26$$

Since relationship 26 must hold for arbitary (a), it follows that,

$$\boxed{|Q| = P\ |W|} \qquad (27$$

Equation 27 represents a set of (n + 1) linear constraints on the (m + 1) ordinates of the weighting function (W) required by 20a.

Evaluation of $E(\epsilon_m^2)$

Let

$$E(N_jN_k) = \sigma^2\rho\left[(j-k)\Delta t\right] \qquad (28$$

be the correlation function of N(t) and

$$E(M_jM_k) = \gamma^2\, r\left[(j-k)\Delta t\right] \qquad (29$$

be the correlation function of M(t) where

$$\rho(0) = r(0) = 1 \qquad (30$$

and let

$$E(M_wN_v) = 0 \qquad (31$$

for all w and v. Then from equations 7, 15, 19 and 25

$$E(\epsilon_m^2) = E\left[M^0 - \sum_{j=0}^{m} W_j(N_{m-j} + M_{m-j})\right]^2 . \qquad (32$$

Let

$$V_{j+1,\,k+1} = \gamma^2 r\left[(j-k)\Delta t\right] + \sigma^2\rho\left[(j-k)\Delta t\right]$$
$$j = 0,\ 1,\ 2,\ \ldots m \text{ and}$$
$$k = 0,\ 1,\ 2,\ \ldots m \qquad (33$$

and let V be the (m+1) x (m+1) matrix whose elements are $V_{j+1,\,k+1}$. Note that V is a symmetric matrix as is its inverse due to the assumption that the random components are stationary.

Define

$$(\beta) = (\beta_1,\ \beta_2,\ \ldots\ \beta_m) \qquad (34$$

$$\beta_j = E\,(M^0\, M_{m-j}) \qquad (35$$

$$L^2 = E\left[(M^0)^2\right] \qquad (36$$

Then L is not a function of the weighting sequence. Substituting 34, 35, and 36 into 32, one obtains

$$E\,(\epsilon^2_m) = (W)\; V\; |W| - 2\,(\beta)\;\; |W| + L^2. \qquad (37$$

When

$$S^*(m\Delta t) = \frac{d^K}{dt^K} S(t) \;\Big|_{t\,=\,(m+\alpha)\Delta t} \qquad (38$$

then

$$Q_k = \frac{d^K}{dt^K} P_k(t) \;\Big|_{t\,=\,(m+\alpha)\Delta t}$$

and

$$L^2 = (-1)^K \gamma^2 \left[\frac{d^{2K}}{dt^{2K}}\; r(t) \;\Big|_{t=0}\right] \qquad (40$$

and

$$\beta_j = \gamma^2 \left[\frac{d^K}{dt^K}\; r(t) \;\Big|_{t\,=\,(j+\alpha)\Delta t}\right] \qquad (41$$

Minimization of $E(\epsilon^2_m)$ with respect to (W)

It is required to minimize $E(\epsilon^2_m) = (W)\; V\,|W| - 2(\beta)\;|W| + L^2$ with respect to (W), subject to the constraints $|Q| = P\;|W|$.

Let

$$|g| = (W)\;\; V\;|W| - 2(\beta)\;|W| + L^2 - 2\left[(Q) - (W)\;P'\right]|\lambda| \qquad (42$$

where $2\;|\lambda| = \begin{vmatrix} 2\lambda_0 \\ \cdot\cdot \\ \cdot\cdot \\ 2\lambda_n \end{vmatrix}$ is a column matrix of (n+1) Lagrangian multipliers. Then the minimum $E(\epsilon^2_m)$ is given by $\frac{\partial g}{\partial w} = 0 = 2V\,|W| - 2|\beta| - 2P'\,|\lambda| = 0$, so that

$$V\,|W| = |\beta| + P'\,|\lambda|\,. \qquad (43$$

Using the restraint equations $|Q| = P\;|W|$ the unknown Lagrangian multipliers are eliminated and the equations solved for |W| giving,

$$|W| = V^{-1}\,|\beta| + V^{-1}P'\left[PV^{-1}P'\right]^{-1}\;|Q| - V^{-1}P'\,(PV^{-1}P')^{-1}\,PV^{-1}\,|\beta| \qquad (44$$

where the superscript -1 indicates the inverse of the matrix. It is assumed that the matrices V and $\left[PV^{-1}P'\right]$ are nonsingular so that their inverses are uniquely defined.

Interpretation of the Weighting Sequence as a Sliding Arc Technique

The output of the digital filter at time $t = m\Delta t$ is given by

$$e^*_m = \sum_{k=0}^{m} W_k e_{m-k} \qquad (45$$

Suppose one requires the output in real time at $t = (m+u)\Delta t$ where u = 1, 2, 3, . . . Then one may write

$$e^*_{m+u} = \sum_{k=0}^{m} W_{k,\,u}\; e_{m+u-k} \qquad (4$$

where the e_{m+u-k} represent the m+1 input values from time $t = (m+u)\Delta t$ to $u\Delta t$. The weighting sequences are now interpreted as a set of sliding weights such that $W_{0,\,u}$ weights the most recent data point in time and $W_{m,\,u}$ weights the last input which occurs $m\Delta t$ previous to the most recent data point. For the general class of input functions P(t), the weighting sequence $W_{k,\,u}$ is time varying, that is it changes for each value of u. The optimum weighting sequence is given by

$$|W_u| = V^{-1}\,|\beta| + V^{-1}P'_u\left[P_u V^{-1} P'_u\right]^{-1}\;|Q_u| - V^{-1}P'_u\left[P_u V^{-1} P'_u\right] P_u V^{-1}\;\;|\beta|\,. \qquad (47$$

Notice that the only changes required are the substitution of P_u for P and Q_u for Q. The other matrices do not change since the random components of the input are assumed to be stationary.

Time Invariant Weighting Sequence

For a certain class of input functions P(t) the weighting sequences have the property that $W_{k,\,u} =$

W_k where W_k is given by equation 44, u = 1, 2, . . . , and k = 0, 1, 2, . . . m. Let H define this class of functions. Then H is specified by the following properties:

a) H is an n + 1 dimensional linear vector space with basis $P_k(t)$, k = 0, 1, 2, . . . n.

b) A translation in time is linear in H, i.e.g.,

$$\sum_{k=0}^{n} a_k P_k(t+\tau) = \sum_{k=0}^{n} a'_k(\tau) P_k(t) \quad \text{for all } \tau. \qquad (48$$

Examples of members of this class are arbitrary polynominals of degree n, exponential functions of the form exp(c+jd) t, sums of the form (a sin ωt + b cos ωt) and products of the above functions. These combinations can be summarized by stating that the complete set of solutions to a homogeneous linear differential equation with constant coefficients of order n+1 belongs to H.

As an example if $P(t) = a t^2$ then P(t) does not belong to H since $P(t+T) = a(t^2 + 2Tt + T^2)$ is of a different form. One may derive an optimum filter which will predict for the function $P(t) = at^2$ and which will require only one equation of constraint. However the weighting sequence of the filter will not be time invariant. If one derived a filter for the more general function $P(t) = a_2t^2 + a_1t + a_0$ then the weighting sequence would be time invariant but would require three equations of constraint and have a larger mean square error of prediction, The filter designed to predict for the general quadratic would also predict correctly for the function $P(t) = a t^2$.

A simple example of the time invariant sliding arc property of the weighting sequence will be shown.

Let $P(t) = a_0 e^{(c+j\omega)t} = a_0 P_0(t)$, and the desired output be $\frac{d}{dt} P(t) \Big|_{t = (m)\Delta t}$, where m = 4.

Let $q = e^{(c+j\omega)\Delta t}$, and M(t) = 0 $\therefore |\beta| = |0|$, and V = I. Then $P = \left[q^4, q^3, q^2, q^1, 1\right]$, $P' = \begin{vmatrix} q^4 \\ q^3 \\ q^2 \\ q^1 \\ q^0 \end{vmatrix}$,

$$\left[P' P\right]^{-1} = \frac{1}{\sum_{j=0}^{4} q^{2j}},$$

$$Q = \frac{d}{dt} P_0(t) \Big|_{t = (m)\Delta t} = (c+j\omega) q^4,$$

$$W = P' \left[P' P\right]^{-1} Q, \text{ and}$$

$$W_K = \frac{(c+j\omega) q^4}{\sum_{j=0}^{4} q^{2j}} q^{4-K}, \text{ so that}$$

$$E(e^*_4) = \sum_{K=0}^{4} W_K a_0 q^{4-K} = a_0(c+j\omega)q^4,$$

$$E\left(e^*_{4+u}\right) = \sum_{K=0}^{4} W_K a_0 q^{4+u-K}$$

$$E(e^*_{4+u}) = \frac{a_0 (c+j\omega) q^4}{\sum_{j=0}^{4} q^{2j}} \left[q^4 q^{4+u} + q^3 q^{3+u} \ldots q^0 q^u\right],$$

$E(e^*_{4+u}) = a_0(c+j\omega) q^{u+4}$ which is the derivative at the end of the prediction interval whose most recent point is (4+u)Δt.

Special Cases

Let M(t) = 0; then $\beta_j = 0$ for all j and equation 47 becomes

$$\left|W_u\right| = V^{-1} P'_u \left[P_u V^{-1} P'_u\right]^{-1} \left|Q_u\right|. \qquad (49$$

Further, if the noise is uncorrelated so that

$$e\left[(j-K)\Delta t\right] = \delta_{jK} \qquad (50$$

then $V^{-1} = \sigma^2 I$, where I is the identity matrix. Then

$$W_u = P'_u \left[P_u P'_u\right]^{-1} \left|Q_u\right| \qquad (51$$

Finally, if the functions $P_k(t)$ are orthogonal, e.g., satisfy the condition

$$\sum_{j=0}^{m} P_k(j+u)\, P_L(j+u) = \delta_{kL} \quad . \qquad (52$$

$$k, L = 0,\ 1,\ 2,\ \ldots n$$

Then $P_u P'_u = I$ and equation 47 becomes

$$\left|W_u\right| = P'_u \left|Q_u\right| \qquad (53$$

Modified Wiener Model

Let $P(t) \equiv 0$ and $S(t) \equiv M(t)$ then

$$e^*(m\Delta t) = \sum_{j=0}^{m} W_j \left[M_{m-j} + N_{m-j}\right],$$

$$S^*(m\Delta t) = \int_{-\infty}^{+\infty} k'(\tau)\ M(m\Delta t - \tau)\, d\tau \quad \text{and}$$

$$\epsilon_m = e^*(m\Delta t) - S^*(m\Delta t) \qquad E(\epsilon_m) = 0 .$$

Equation 37 gives the value of $E(\epsilon_m^2)$.

$$\frac{\partial E(\epsilon_m^2)}{\partial W} = V\left|W\right| - \left|\beta\right| = 0 \therefore \left|W\right| = V^{-1}\left|\beta\right| \qquad (54$$

Conclusion

A general relationship for the optimum weighting sequence has been derived such that the mean square error of prediction is a minimum. The model for the input signal contains a nonstationary component which is an arbitrary linear combination of n+1 known functions plus a stationary random component and stationary noise. The condition for which the weighting function can be utilized as a time invariant sliding arc has been stated and an example of this technique applied.

Appendix I: Extension of the Digital Filter to a Continuous, Piecewise Analytic Filter

Let the input to the digital filter consist of a polynomial of degree n and white noise. Then

$$M(t) = 0 \qquad (1a$$

and

$$\rho\left[(j-k)\Delta t\right] = \delta_{jk} . \qquad (2a$$

The input polynomial is defined by

$$P(j\Delta t) = \sum_{k=0}^{n} a_k P_k(j\Delta t) . \qquad (3a$$

The $P_k(j\Delta t)$ are selected to be orthogonal over the interval $j = 0,\ 1,\ 2,\ \ldots\ m$. That is $P_k(j\Delta t)$ satisfies the relationship

$$\sum_{j=0}^{m} P_k(j\Delta t)\, P_h(j\Delta t) = \delta_{k,h} S(k,\ m+1) . \qquad (4a$$

The first few polynomials are given by the relationships

$$P_k(j\Delta t) = \frac{\epsilon_k(j\Delta t)}{\left[S(k, m+1)\right]}\ 1/2 \qquad (5a$$

where

$$\epsilon_0(j\Delta t) = 1$$

$$\epsilon_1(j\Delta t) = \Delta t\left((j+1) - \frac{m+2}{2}\right) \qquad (6a$$

$$\epsilon_2(j\Delta t) = \left[\epsilon_1^2(j\Delta t) - \frac{(m+1)^2 - 1}{12}\right]$$

and $S(k, m+1) = \sum_{j=0}^{m} \epsilon_k^2(j\Delta t)$.

Higher order polynomials can be obtained from reference 6.

Let the desired output be

$$S^*(m\Delta t) = \frac{d^L}{dt^L} P(t)\Big|_{t = (m+\alpha)\Delta t} . \qquad (7a$$

Then

$$Q_k = \frac{d^L}{dt^L} P_k(t)\Big|_{t = (m+\alpha)\Delta t}$$

$$\equiv P_k^{(L)}\left[(m+\alpha)\Delta t\right] . \qquad (8a$$

Then the weighting sequence by equation 53 is (for u equal to zero)

$$\left|W\right| = P'\ \left|Q\right| \qquad (9a$$

so that

$$W_v = \sum_{k=L}^{n} Q_k P_k\left[(m-v)\Delta t\right] , \qquad (10a$$

where v = 0, 1, 2, . . m. Thus for n = 1 and L = 0 one has a linear input and an m + 1 point least squares curve fit. The output is the predicted value of the input where $\alpha\Delta t$ is the prediction interval.

Then

$$W_v = \frac{1}{m+1} + \frac{\epsilon_1 \left[(m-v)\Delta t\right] \; \epsilon_1 \left[(m+\alpha)\Delta t\right]}{S\,(1,\ m+1)}\ , \qquad (11a$$

and using equation 6a one has

$$W_v = \frac{1}{m+1} + \frac{(\Delta t)^2 \left[(m+1-v) - \frac{m+2}{2}\right]\left[(m+\alpha+1) - \frac{m+2}{2}\right]}{S\,(1, m+1)}\ . \qquad (12a$$

If now the variable $\alpha+\Delta\alpha$ is substituted for α where $0 \leqq \Delta\alpha \leqq 1$, then W_v is a continuous function of $\Delta\alpha$. This substitution is equivalent to a continuous minimum variance prediction over the interval between the samples based on the previous knowledge of the last m + 1 sample points. When an additional data point is sampled, then the constant corresponding to $\Delta\alpha = 0$ is used. That is

$$W_v\,(\Delta\alpha) = W_v\,(\Delta\alpha + \lambda) \qquad (13a$$

where $\lambda = 0, \pm 1, \pm 2, \ldots$. The W_v are cyclic with a period of unity.

Upon making the above substitution in equation 12a) one has

$$W_v = \frac{1}{m+1} + \frac{\left[\frac{m}{2} - v\right]\left[\frac{m}{2} + \left[\alpha+\Delta\alpha\right]\right]}{S\,(1,\ m+1)}\ . \qquad (14a$$

(Note: Δt is taken equal to unity.)

For m = 2,

$$W_v = \frac{1}{3} + \frac{1}{2}\left[1-v\right]\left[1+\alpha+\Delta\alpha\right] \qquad (15a$$

Since $\epsilon_1(0) = -1$, $\epsilon_1(1) = 0$, $\epsilon_1(2) = 1$ then $S(1, 3) = 2$, from which one obtains,

$$W_0 = \frac{5}{6} + \frac{\alpha+\Delta\alpha}{2}$$
$$W_1 = \frac{1}{3} \qquad (16a$$
$$W_2 = -\frac{1}{6} - \frac{1}{2}\,(\alpha+\Delta\alpha)\ .$$

The above solution checks with the functions $u_0(t)$, and $u_1(t)$, and $u_2(t)$ presented by Lees for this same model.

As a second example consider n = 1, L = 0, m = 3. Then

$$W_v = \frac{1}{4} + \frac{\left[\frac{3}{2} - v\right]\left[\frac{3}{2} + (\alpha+\Delta\alpha)\right]}{S(1, 4)}\ . \qquad (17a$$

Since $\epsilon_1(0) = -3/2$, $\epsilon_1(1) = -1/2$, $\epsilon_1(2) = +1/2$, $\epsilon_1(3) = 3/2$, then $S(1, 4) = \sum_{j=0}^{3} \epsilon_1(j) = \frac{20}{4}$. One evaluates

$$W_v = \frac{1}{4} + \frac{\left[3 - 2v\right]\left[3 + 2\,(\alpha+\Delta\alpha)\right]}{20} \qquad (18a$$

where v = 0, 1, 2, and 3, and obtains

$$W_0 = \frac{7 + 3\,(\alpha+\Delta\alpha)}{10} = u_0(\Delta\alpha)$$
$$W_1 = \frac{4 + 1\,(\alpha+\Delta\alpha)}{10} = u_1(\Delta\alpha)$$
$$W_2 = \frac{+1 - 1\,(\alpha+\Delta\alpha)}{10} = u_2(\Delta\alpha) \qquad (19a$$
$$W_3 = -\frac{(2 + 3\,(\alpha+\Delta\alpha)}{10} = u_3(\Delta\alpha)\ .$$

These results check with the functions obtained by Lees.

Approaching the solution from the point of view of the continuous extension of α to $\alpha+\Delta\alpha$, certain properties become apparent for the above input model as follows:

a) The mean square error output is smallest at the times corresponding to the sampling of a new data point and increases monotonically until the next data point. At the next data point the mean square error returns to the previous value. Thus

$$\sigma^2(\Delta\alpha) = \sigma^2(\Delta\alpha+\lambda) \qquad (20a$$

for $\lambda = 0, \pm1, \pm2, \ldots$. That is, the mean square error is periodic.

b) The smallest possible value of σ^2 is given when

$$\alpha + \Delta\alpha = -\left[\frac{m+1}{2}\right] \qquad (21a$$

c) The values of σ^2 and $W_v(\Delta\alpha)$ are functions of variables α and $\Delta\alpha$ in the form $\alpha-\Delta\alpha$ only.

d) For an L^{th} derivative output $W_v(\alpha+\Delta\alpha)$ is a polynomial of degree n - L, $(n \geq L)$ in $(\alpha+\Delta\alpha)$ and a polynomial of degree n in v.

If the analytic extension of α to $\alpha+\Delta\alpha$ is taken as a solution to the optimum continuous filter for discrete data then one may use equations 44 and 47 as the solution to this problem. One need only substitute $\alpha+\Delta\alpha$ for α in the Q and B matricies since, if one assumes stationary noise, the other matrixes are not affected. It is not claimed by the author that the extension is proved herein. From an intuitive point of view however since one has no information other than at the sampled points, it would seem that the best one could do between samples is to extrapolate in an optimum manner using the minimum variance filter.

The fact that this procedure checks out in the case tested lends qualitative weight to use of the same principle in the extended model. For the class of inputs belonging to H, the weighting function will be time invariant and the periodic properties of the mean square error will be the same as previously discussed. For the more general input function not in H, the filter will be a time varying function. Thus as each sample is presented to the filter, a new set of m + 1 functions $W_v(\alpha+\Delta\alpha)$ will be required, as given by equation 47.

The mean square error will be smallest at the sampled points and increase monotonically till the next sample point. At the next sample point the mean square error changes discontinuously to the smallest value for the next interval.

References

1. Norbert Wiener, "Extropolation, Interpolation and Smoothing of Stationary Time Series with Engineering Application", Cambridge Technology Press of M.I.T., 1949.

2. The National Military Establishment Research and Development Board, "Data Smoothing Prediction in Fire Control, Appendix B" Report Series No. 13, MCG 12/1, 15 August 1948.

3. Zadeh and Ragazzini, "An Extension of Wiener's Theory of Prediction", Journal of Applied Physics, Volume 21, July 1950.

4. Lees, A. B., "Interpolation and Extrapolation of Sampled Data", IRE Transactions on Information Theory, Volume IT-2, March 1956.

5. Aitken, A. C., "On Least Squares and Linear Combination of Observations", Proceedings of Royal Society of Edinburgh, Volume 55, 1934-35, pp. 42-48.

6. Anderson, R. L. and Hauseman, E. E. "Tables of Orthogonal Polynomial Values Extended to N = 104", Research Bulletin 297, April 1942, Ames, Iowa (Iowa State College).

A NEW INTERPRETATION OF INFORMATION RATE

by

J. L. KELLY, JR.
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

(Reprinted from B.S.T.J., July 1956)

ABSTRACT

If the input symbols to a communication channel represent the outcomes of a chance event on which bets are available at odds consistent with their probabilities (i.e., "fair" odds), a gambler can use the knowledge given him by the received symbols to cause his money to grow exponentially. The maximum exponential rate of growth of the gambler's capital is equal to the rate of transmission of information over the channel. This result is generalized to include the case of arbitrary odds.

Thus we find a situation in which the transmission rate has significance even though no coding is contemplated. Previously this quantity was given significance only by a theorem of Shannon's which asserted that, with suitable encoding, binary digits could be transmitted over the channel at this rate with an arbitrarily small probability of error.

Introduction

Shannon defines the rate of transmission over a noisy communication channel in terms of various probabilities.[1] This definition is given significance by a theorem which asserts that binary digits may be encoded and transmitted over the channel at this rate with arbitrarily small probability of error. Many workers in the field of communication theory have felt a desire to attach significance to the rate of transmission in cases where no coding was contemplated. Some have even proceeded on the assumption that such a significance did, in fact, exist. For example, in systems where no coding was desirable or even possible (such as radar), detectors have been designed by the criterion of maximum transmission rate or, what is the same thing, minimum equivocation. Without further analysis such a procedure is unjustified.

The problem then remains of attaching a value measure to a communication system in which errors are being made at a non-negligible rate, i.e., where optimum coding is not being used. In its most general formulation this problem seems to have but one solution. A cost function must be defined on pairs of symbols which tell how bad it is to receive a certain symbol when a specified signal is transmitted. Furthermore, this cost function must be such that its expected value has significance, i.e., a system must be preferable to another if its average cost is less. The utility theory of Von Neumann[2] shows us one way to obtain such a cost function. Generally this cost function would depend on things external to the system and not on the probabilities which describe the system, so that its average value could not be identified with the rate as defined by Shannon.

The cost function approach is, of course, not limited to studies of communication systems, but can actually be used to analyze nearly any branch of human endeavor. The author believes that it is too general to shed any light on the specific problems of communication theory. The distinguishing feature of a communication system is that the ultimate receiver (thought of here as a person) is in a position to profit from any knowledge of the input symbols or even from a better estimate of their probabilities. A cost function, if it is supposed to apply to a communication system, must somehow reflect this feature. The point here is that an arbitrary combination of a statistical transducer (i.e., a channel) and a cost function does not necessarily constitute a communication system. In fact(not knowing the exact definition of a communication system on which the above statements are tacitly based) the author would not know how to test such an arbitrary combination to see if it were a communication system.

What can be done, however, is to take some real-life situation which seems to possess the essential features of a communication problem, and to analyze it without the introduction of an arbitrary cost function. The situation which will be chosen here is one in which a gambler uses knowledge of the received symbols of a communication channel in order to make profitable bets on the transmitted symbols.

The Gambler With A Private Wire

Let us consider a communication channel which is used to transmit the results of a chance situation before those results become common knowledge, so that a gambler may still place bets at the original odds. Consider first the case of a noiseless binary channel, which might be used, for example, to transmit the results of a series of baseball games between two equally matched teams. The gambler could obtain even money bets even though he already knew the result of each game. The amount of money he could make would depend only on how much he chose to bet. How much would he bet? Probably all he had since he would win with certainty. In this case his capital would grow exponentially and after n bets he would have 2^n times his original bankroll. This exponential growth of capital is not uncommon in economics. In fact, if the binary digits in the above channel

were arriving at the rate of one per week, the sequence of bets would have the value of an investment paying 100 per cent interest per week compounded weekly. We will make use of a quantity G called the exponential rate of growth of the gambler's capital, where

$$G = \lim_{n\to\infty} \frac{1}{n} \log \frac{V_n}{V_o}$$

where V_n is the gambler's capital after n bets, V_o is his starting capital, and the logarithm is to the base two. In the above example G = 1.

Consider the case now of a noisy binary channel, where each transmitted symbol has probability, p, or error and q of correct transmission. Now the gambler could still bet his entire capital each time, and, in fact, this would maximize the expected value of his capital, $\langle V_n \rangle$, which in this case would be given by

$$\langle V_n \rangle = (2q)^n V_o \ .$$

This would be little comfort, however, since when n was large he would probably be broke and, in fact, would be broke with probability one if he continued indefinitely. Let us, instead, assume that he bets a fraction, ℓ, of his capital each time. Then

$$V_n = (1+\ell)^W (1-\ell)^L V_o$$

where W and L are the number of wins and losses in the n bets. Then

$$G = \lim_{n\to\infty} \left[\frac{W}{n} \log(1+\ell) + \frac{L}{n} \log (1-\ell) \right]$$

$$= q \log(1+\ell) + p \log(1-\ell) \quad \text{with probability one.}$$

Let us maximize G with respect to ℓ. The maximum value with respect to the Y_i of a quantity of the form $Z = \Sigma X_i \log Y_i$, subject to the constraint $\Sigma Y_i = Y$, is obtained by putting

$$Y_i = \frac{Y}{X} X_i,$$

where $X = \Sigma X_i$. This may be shown directly from the convexity of the logarithm.

Thus we put

$$(1+\ell) = 2q$$

$$(1-\ell) = 2p$$

and

$$G_{max} = 1 + p \log p + q \log q$$

$$= R$$

which is the rate of transmission as defined by Shannon.

One might still argue that the gambler should bet all his money (make $\ell = 1$) in order to maximize his expected win after n times. It is surely true that if the game were to be stopped after n bets the answer to this question would depend on the relative values (to the gambler) of being broke or possessing a fortune. If we compare the fates of two gamblers, however, playing a non-terminating game, the one which uses the value ℓ found above will, with probability one, eventually get ahead and stay ahead of one using any other ℓ. At any rate, we will assume that the gambler will always bet so as to maximize G.

The General Case

Let us now consider the case in which the channel has several input symbols, not necessarily equally likely, which represent the outcome of chance events. We will use the following notation:

- $p(s)$ — the probability that the transmitted symbol is the s'th one.
- $p(r/s)$ — the conditional probability that the received symbol is the r'th on the hypothesis that the transmitted symbol is the s'th one.
- $p(s,r)$ — the joint probability of the s'th transmitted and r'th received symbol.
- $q(r)$ — received symbol probability.
- $q(s/r)$ — conditional probability of transmitted symbol on hypothesis of received symbol.
- α_s — the odds paid on the occurrence of the s'th transmitted symbol, i.e., α_s is the number of dollars returned for a one-dollar bet (including that one dollar).
- $a(s/r)$ — the fraction of the gambler's capital that he decides to bet on the occurrence of the s'th transmitted symbol after observing the r'th received symbol.

Only the case of independent transmitted symbols and noise will be considered. We will consider first the case of "fair" odds, i.e.,

$$\alpha_s = \frac{1}{p(s)} \ .$$

In any sort of parimutuel betting there is a tendency for the odds to be fair (ignoring the "track take"). To see this first note that if there is no "track take"

$$\Sigma \frac{1}{\alpha_s} = 1$$

since all the money collected is paid out to the winner. Next note that if

$$\alpha_s > \frac{1}{p(s)}$$

for some s a bettor could insure a profit by making repeated bets on the s'th outcome. The extra

betting which would result would lower α_s. The same feedback mechanism probably takes place in more complicated betting situations, such as stock market speculation.

There is no loss in generality in assuming that

$$\sum_s a(s/r) = 1$$

i.e., the gambler bets his total capital regardless of the received symbol. Since

$$\sum_s \frac{1}{\alpha_s} = 1$$

he can effectively hold back money by placing canceling bets. Now

$$V_N = \prod_{r,s} [a(s/r)\,\alpha_s]^{W_{sr}}\, V_o$$

where W_{sr} is the number of times that the transmitted symbol is s and the received symbol is r.

$$\begin{aligned} \operatorname{Log} \frac{V_n}{V_o} &= \sum_{rs} W_{sr} \log \alpha_s a(s/r) \\ G = \lim_{n\to\infty} \frac{1}{N} \log \frac{V_n}{V_o} &= \sum_{rs} p(s,r) \log \alpha_s a(s/r) \end{aligned} \qquad (1)$$

with probability one. Since

$$\alpha_s = \frac{1}{p(s)}$$

here

$$\begin{aligned} G &= \sum_{rs} p(s,r) \log \frac{a(s/r)}{p(s)} \\ &= \sum_{rs} p(s,r) \log a(s/r) + H(X) \end{aligned}$$

where H(X) is the source rate as defined by Shannon. The first term is maximized by putting

$$a(s/r) = \frac{p(s,r)}{\sum_k p(k,r)} = \frac{p(s,r)}{q(r)} = q(s/r).$$

Then $G_{max} = H(X) - H(X/Y)$, which is the rate of transmission defined by Shannon.

When the Odds are Not Fair

Consider the case where there is no track take, i.e.,

$$\sum_s \frac{1}{\alpha_s} = 1$$

but where α_s is not necessarily

$$\frac{1}{p(s)}.$$

It is still permissible to set $\sum_s a(s/r) = 1$ since the gambler can effectively hold back any amount of money by betting it in proportion to the $1/\alpha_s$. Equation (1) now can be written

$$G = \sum_{rs} p(s,r) \log a(s/r) + \sum_s p(s) \log \alpha_s.$$

G is still maximized by placing $a(s/r) = q(s/r)$ and

$$\begin{aligned} G_{max} &= -H(X/Y) + \sum_s p(s) \log \alpha_s \\ &= H(\alpha) - H(X/Y) \end{aligned}$$

where

$$H(\alpha) = \sum_s p(s) \log \alpha_s .$$

Several interesting facts emerge here

(a) In this case G is maximized as before by putting $a(s/r) = q(s/r)$.

That is, the gambler ignores the posted odds in placing his bets!

(b) Since the minimum value of $H(\alpha)$ subject to

$$\sum_s \frac{1}{\alpha_s} = 1$$

obtains when

$$\alpha_s = \frac{1}{p(s)}$$

and $H(X) = H(\alpha)$, any deviation from fair odds helps the gambler.

(c) Since the gambler's exponential gain would $H(\alpha) - H(X)$ if he had no inside information, we can interpret $R = H(X) - H(X/Y)$ as the increase of G_{max} due to the communication channel. When there is no channel, i.e., $H(X/Y) = H(X)$, G_{max} is minimized (at zero) by setting

$$\alpha_s = \frac{1}{p_s} .$$

This gives further meaning to the concept "fair odds."

When There is a "Track Take"

In the case there is a "track take" the situation is more complicated. It can no longer be assumed that $\sum_s a(s/r) = 1$. The gambler cannot make canceling bets since he loses a percentage to the track. Let $b_r = 1 - \sum_s a(s/r)$, i.e., the fraction not bet when the received symbol is the r'th one. Then the quantity to be maximized is

$$G = \sum_{rs} p(s,r) \log b_r + \alpha_s a(s/r) \quad , \qquad (2)$$

subject to the constraints

$$b_r + \sum_s a(s/r) = 1.$$

In maximizing (2) it is sufficient to maximize the terms involving a particular value of r and to do this separately for each value of r since both in (2) and in the associated constraints, terms involving different r's are independent. That is, we must maximize terms of the type

$$G_r = q(r) \sum_s q(s/r) \log b_r + \alpha_s a(s/r)$$

subject to the constraint

$$b_r + \sum_s a(s/r) = 1 .$$

Actually, each of these terms is the same form as that of the gambler's exponential gain where there is no channel

$$G = \sum_s p(s) \log b + \alpha_s a(s) . \qquad (3)$$

We will maximize (3) and interpret the results either as a typical term in the general problem or as the total exponential gain in the case of no communication channel. Let us designate by λ the set of indices, s, for which $a(s) > 0$, and by λ' the set for which $a(s) = 0$. Now at the desired maximum

$$\frac{\partial G}{\partial a(s)} = \frac{p(s)\alpha_s}{b + a(s)\alpha_s} \log e = k \quad \text{for } s\varepsilon\lambda$$

$$\frac{\partial G}{\partial b} = \sum_s \frac{p(s)}{b + a(s)\alpha_s} \log e = k$$

$$\frac{\partial G}{\partial a(s)} = \frac{p(s)\alpha_s}{b} \log e \leq k \quad \text{for } s\varepsilon\lambda'$$

where k is a constant. The equations yield

$$k = \log e, \quad b = \frac{1 - p}{1 - \sigma}$$

$$a(s) = p(s) - \frac{b}{\alpha_s} \quad \text{for } s\varepsilon\lambda$$

where $p = \Sigma_\lambda p(s)$, $\sigma = \Sigma_\lambda (1/\alpha_s)$, and the inequalities yield

$$p(s)\alpha_s \leq b = \frac{1 - p}{1 - \sigma} \quad \text{for } s\varepsilon\lambda'.$$

We will see that the conditions

$$\sigma < 1$$

$$p(s)\alpha_s > \frac{1 - p}{1 - \sigma} \quad s\varepsilon\lambda$$

$$p(s)\alpha_s \leq \frac{1 - p}{1 - \sigma} \quad \text{for } s\varepsilon\lambda$$

completely determine λ.

If we permute indices so that

$$p(s)\alpha_s \geq p(s+1)\alpha_{s+1}$$

then λ must consist of all $s \leq t$ where t is a positive integer or zero. Consider how the fraction

$$F_t = \frac{1 - p_t}{1 - \sigma_t}$$

varies with t, where

$$p_t = \sum_1^1 p(s), \quad \sigma_t = \sum_1^1 \frac{1}{\alpha_s} ; \quad F_o = 1.$$

Now if $p(1)\alpha_1 < 1$, F_t increases with t until $\sigma_t \geq 1$. In this case $t = 0$ satisfies the desired conditions and λ is empty. If $p(1)\alpha_1 > 1$ F_t decreases with t until $p(t+1)\alpha_{t+1} < F_t$ or $\sigma_t \geq 1$. If the former occurs, i.e., $p(t+1)\alpha_{t+1} < F_t$, then $F_{t+1} > F_t$ and the fraction increases until $\sigma_t \geq 1$. In any case the desired value of t is the one which gives F_t its minimum positive value, or if there is more than one such value of t, the smallest. The maximizing process may be summed up as follows:

(a) Permute indices so that $p(s)\alpha_s \geq p(s+1)\alpha_{s+1}$

(b) Set b equal to the minimum positive value of

$$\frac{1-p_t}{1-\sigma_t} \quad \text{where} \quad p_t = \sum_1^1 p(s), \; \sigma_t = \sum_1^1 \frac{1}{\alpha_s}$$

(c) Set $a(s) = p(s) - b/\alpha_s$ or zero, whichever is larger. (The a(s) will sum to 1-b.)

The desired maximum G will then be

$$G_{max} = \sum_1^1 p(s) \log p(s)\alpha_s + (1-p_t) \log \frac{1-p_t}{1-\sigma_t}$$

where t is the smallest index which gives

$$\frac{1-p_t}{1-\sigma_t}$$

its minimum positive value.

It should be noted that if $p(s)\alpha_s < 1$ for all s no bets are placed, but if the largest $p(s)\alpha_s > 1$ some bets might be made for which $p(s)\alpha_s < 1$, i.e., the expected gain is negative. This violates the criterion of the classical gambler who never bets on such an event.

Conclusion

The gambler introduced here follows an essentially different criterion from the classical gambler. At every bet he maximizes the expected value of the logarithm of his capital. The reason has nothing to do with the value function which he attached to his money, but merely with

the fact that it is the logarithm which is additive in repeated bets and to which the law of large numbers applies. Suppose the situation were different; for example, suppose the gambler's wife allowed him to bet one dollar each week but not to reinvest his winnings. He should then maximize his expectation (expected value of capital) on each bet. He would bet all his available capital (one dollar) on the event yielding the highest expectation. With probability one he would get ahead of anyone dividing his money differently.

It should be noted that we have only shown that our gambler's capital will surpass, with probability one, that of any gambler apportioning his money differently from ours but still in a fixed way for each received symbol, independent of time or past events. Theorems remain to be proved showing in what sense, if any, our strategy is superior to others involving a(s/r) which are not constant.

Although the model adopted here is drawn from the real-life situation of gambling it is possible that it could apply to certain other economic situations. The essential requirements for the validity of the theory are the possibility of reinvestment of profits and the ability to control or vary the amount of money invested or bet in different categories. The "channel" of the theory might correspond to a real communication channel or simply to the totality of inside information available to the investor.

Let us summarize briefly the results of this paper. If a gambler places bets on the input symbol to a communication channel and bets his money in the same proportion each time a particular symbol is received his, capital will grow (or shrink) exponentially. If the odds are consistent with the probabilities of occurrence of the transmitted symbols (i.e., equal to their reciprocals), the maximum value of this exponential rate of growth will be equal to the rate of transmission of information. If the odds are not fair, i.e., not consistent with the transmitted symbol probabilities but consistent with some other set of probabilities, the maximum exponential rate of growth will be larger than it would have been with no channel by an amount equal to the rate of transmission of information. In case there is a "track take" similar results are obtained, but the formulae involved are more complex and have less direct information theoretic interpretations.

Acknowledgements

I am indebted to R. E. Graham and C. E. Shannon for their assistance in the preparation of this paper.

References

1. C. E. Shannon, A Mathematical Theory of Communication, B.S.T.J., 27, pp.379-423, 623-656, Oct., 1948.

2. Von Neumann and Morgenstein, Theory of Games and Economic Behavior, Princeton Univ. Press, 2nd Edition, 1947.

AN OUTLINE OF A PURELY PHENOMENOLOGICAL THEORY OF STATISTICAL THERMODYNAMICS : I. CANONICAL ENSEMBLES

by

Benoit MANDELBROT
Faculté des Sciences de l'Université de Genève
Genève, Suisse

Summary. "Boltzmann's problem" of statistical thermodynamics, is that of eliminating the paradoxical incompatibility of structure, existing between the irreversibility of the classical phenomenological thermodynamics, and the reversibility of any purely kinetic model, one could ever think of for these phenomena. One finds that, in order to construct kinetic "analogs" to the laws of phenomenological thermodynamics, the dynamics of large assemblies of molecules (Liouville theorem, etc. ...) must be completed by some hypotheses of randomness. Once established, this randomness can be followed up in its development, with no new conceptual paradox (although with great technical difficulty : there is a great amount of current work on this topic). But the introduction of randomness still raises entirely uncleared problems. Since, therefore, the kinetic foundations of thermodynamics are not sufficient in the absence of further hypotheses of randomness, are they still quite necessary in the presence of such hypotheses? Or else, could not one "short-circuit" the atoms, by centering upon any elements of randomness, for example those introduced by the process of observation? Our aim is to show (partly after Szilard) that a substantial part of the results, usually obtained through kinetic arguments, could be obtained by postulating from the outset a statistical distribution for the properties of a system, and following up with a purely phenomenological argument. The spirit of the theory is extremely close to that of the conventional (Copenhagen) approach to quantum theory, and the results are quite parallel, although the mathematics is quite different. Randomness is introduced by following the modern statistical theory of the estimation of non directly observable intensive variables of state, such as the temperature. The discussion of the methodological foundations of modern statistics can thus be translated into a full-fledged, and possibly significant, counterpart of the discussion of the kinetic foundations of thermodynamics. Statistics is thus provided with a particularly concrete example for some of its more involved methods; thermodynamics appears clarified in its classical aspects, and is further completed with an apparently new uncertainty relationship. It may also be of interest to the communication engineer to have a unified treatment of the foundations of fluctuation phenomena, and of methods of fighting noise : a discussion of entropy and information, performed in this spirit, will be given in Part II of the paper.

1. Introduction

1.1. The nature of the problem.(+)

The aim of this paper differs in one essential respect from that of most investigations on statistical problems, in communication and elsewhere. This difference should be stated at the outset, in order to avoid certain misunderstandings. Most authors in statistics are concerned with broadly engineering problems, of the improvement of the design of coding and detection procedures, and more generally, of methods of testing hypotheses, and parameter estimation. These problems assume that a sufficiently complete understanding and description of the necessary laws of nature has already been acquired elsewhere. In some cases, the needs of communication have lead to an improvement of the laws of physics. Our problem is precisely the opposite: we wish to improve the knowledge and presentation of precisely those laws of physics, which enter in the above engineering problem of communication, by centering upon the role they play in these problems, and by using the concepts and the mathematics derived for their purposes.

Any degree of success, we could achieve, would be new proof of the fact, which is of course quite familiar, that well-chosen engineering problems often bring out the essentials of a physical situation, in a way that is useful in a far wider context. Such used to be the role of heat engines in thermodynamics. Later on, one tried to use the problems of coding for the same purpose : the starting point of those attempts was the misleadingly simple-looking problem, raised by the fact that the definition, by Shannon[6], of the information involved in communication, is mathematically identical to a classical definition

(+) In order to emphasize the essential similarity of approach, this § was made a paraphrase of the introductory lines of our[1], on the statistical structure of natural languages ; see also [2,3,4,5]

of entropy by Boltzmann. The study of this problem was brought together with that of Maxwell's Demon (we shall see that it is no accident, since these are the only two problems of so-called-statistical thermodynamics, which involve an observer's decision, that is, are actually statistical in J. Neyman's terminology[7]). However, none of the now numerous attempts, to clarify the relationship of information to entropy, is generally felt to have brought much to either communication or thermodynamics. This may be due to the fact that, although information is as much as the coder, or the Demon, need and want to know about certain processes of measurement, (in order to establish their global balance sheet), this much is very little. It is already too little for the communication problem of detection; all the more so, it would be too little to even hope to base upon it a better conceptual understanding of the role of approximate observation (that is : of any observer) in thermodynamics. (Besides, the probability p introduced in Boltzmann's formula $S = k \log p$, is a very queer kind of probability anyway; there is a difficulty there, without counterpart in communication, where one always has a set of prior probabilities for the possible signals, that is, is in the "Bayes" case of statistics). If so, there could be no useful solution to the problem of entropy and information, considered alone. However, the more general problem, of the role of observation and of the observer in thermodynamics, could now be studied in detail, with the help of the statistical theory of estimation (or detection), which has now become a very general methodological model of experience, as inductive behaviour of the physicist in the face of the unknown.[7]

It will be attempted to build a theory of thermodynamics, which will be statistical as well as phenomenological, around the problem of the statistical estimation of state variables; the problem of entropy and information will constitute an application of the theory. (A previous attempt by the author in chap 4 of [3] is now quite obsolete, but the philosophy of models given in other chap. of that reference stands fortified by the present work). Following Kramers, statistical thermodynamics will be referred to as "thermostatistics". Szilard's [8] previous approach is of the greatest relevance for the problem.

One possible conceptual difficulty of this approach will be due to the exclusive use of classical probability concepts, in making clear exactly which phenomena and variables are to be considered as random. When first introduced into statistics by J. Neyman, classical probability encountered "deep rooted habits of thought". Besides, the approach will be very theoretical; but Whitehead may have been right again by asserting that "often in our most theoretical moods we may be closest to the most practical applications." On the contrary, the mathematics will be quite simple, although not usual in information theory studies. This is because we can choose the engineering problems we "invert", to be as simple as possible, which is of course not the case in engineering problems imposed by practical needs.

1.2. The nature of the result

Let us elaborate upon this problem. One aim of communication theory is to find ways and means satisfying certain criteria of quality, by which a signal could be detected through a background of noise. It has been recognized for some time [9,10,11] that this problem is, in principle, simply one of estimating "at best" the emitted signal, S_e, knowing the received signal S_r. For that, S_r is considered to be an observation from a random (because perturbed by noise) population of signals, and S_e is the parameter of this population.

The distribution $p(S_r/S_e)$ is considered to be a part of the engineer's scientific knowledge of nature, that is : at worst, it must be determined by special observation, at best, (if noise is the least compatible with the "structure of matter") it is given by physical laws of more general validity: those of thermodynamics and of quantum theory. We shall show that, conversely, the conventional laws of thermostatistics, from the foundations up, and a few new laws, can be obtained by caracterizing thermal noise as being the "least disturbing for the physicist". A certain amount of imprecision in what is the "least disturbing" will be shown to be allowed by a corresponding familiar imprecision of thermodynamics. The main criterion involved, that of "sufficiency of certain valuations of observation", is not even a variational criterion : it postulates the impossibility of a certain inference, and is an authentic counterpart of the exclusion of certain heat engines by the Carnot principle. Anyway, whichever arbitrary anthropo-centered character may remain in the theory, will be a counterpart of the arbitrary assumptions about molecules, required in the kinetic models, which aim to explain why noise is the least disturbing for the observer which is considered.

Further methodological discussion of the approach will be made in § 4, when the systems sutdied are defined (§2), and the Maxwell Botzmann distribution is derived in several ways (§3). Final discussion will be included in §6.

2. States of a statistical physical system

2.1 Restriction to one-parameter systems

In any truly stochastic physical theory, the relationships between the various state variables of a system are assumed, from the outset, to be ruled by probability laws. The degree of complication of a theory is then determined by the most complicated family of probability distributions considered in that theory. We shall start by a restricted problem, in which the only families to appear will depend upon a single real parameter. The case of several parameters, and the non-parametric case will be considered later; the generality of the approach will appear only then, and it will be seen that one can obtain directly a theory taking account of exchanges of matter and of quantum effects.

2.2 Definitions

Consider a set of methods of measurement, defining a certain level of refinement of the physical analysis. Physical variables can then be of several kinds, distinguished as being, on one side, extensive or intensive; on the other side, as being observable or estimable.

Observable variables are those which can be considered as random variables (r.v.)before measurement (and expressed by capital letters, such as U); and can be actually and directly measured, by single real (scalar continuous) numbers (expressed by lower case letters, such as u). The precision of results of measurement will always be taken as infinite, which means that, following the quantum theory, we shall consider that, by measuring an observable, one actually and physically puts the system in a "state" described by the value which has been found. The measurement may of necessity be infinitely slow. Of course, even then, the results of measurement are usually given only within a margin of possible error, relative to a "truer" value; the latter would be a r.v., depending upon the system itself, whereas the error would be a r.v. depending upon the process of observation. But in fact, there is no criterion for sharing the contributions of the two sources of randomness: at best, the "true" value could be considered as an estimable.

Consider now a system before measurement. Its "state" is a probability distribution function (p.d.f.); it can be considered as a "mixture" or "superposition" of "states" after measurement. If necessary, the state will be called "mixed", as opposed to "pure" states after measurement. This is a relative concept, since a state which is pure with respect to one observable may also be "mixed" with respect to another. A pure and a mixed state are complementary uncompatible descriptions of a system. Measurement thus involves an unpredictable sudden jump from state to state, or rather from being partly in several states into being in one: this has nothing shocking in detection theory.

It will be assumed that the only way, to realize a "microanonical set" of systems, all having equal values for some observable, will be to pick them one by one, after measurement. Thus no "intensive" observables will be considered.

Estimable variables are those parameters, which express the dependence of the probability distributions of the observables, upon the properties of the physical system studied. One must therefore consider populations of systems (the "collectives", in the older terminology), all having the same value for some estimable. When such an equality can be considered as realised by a physical interaction, the estimable is called "intensive". However, much of the theory holds for any estimables.

The obvious example of an observable is the energy U; the corresponding estimable variable is the inverse temperature B (the "observable" character of the energy is an universal axiom of physics, although sometimes hard to justify). A measure of energy is, for example, obtained with a so called "thermometer", by observing the change in its geometrical shape when it absorbs U. On the other side, the uniformity of the B's of a set of systems can be ascertained by letting them a sufficiently long time in thermal contact. (It is not useful to say that fixing B is a kind of measure of it!) Let the probability distribution and eventually density, of U be given by P(u/B) and p(u/B) :

$$P(u/B) = Pr.\left[U \leq u/B\right] \; ; \; p(u/B) = dP/dB$$

Both kinds of variables have been defined within a one-to-one transformation, only. Sometimes, there exists a particularly intrinsic determination of an observable, which is additive, i.e. such that the u of any union of disjoint sets is the sum of the separate u_i. The p(u/B) is then given, knowing the $p_i(u_i/B)$, by the "convolution", iteration of :

$$p(u/B) = \int_0^u p_1\ (u_1/B)\ p_2\ (u - u_1/B)\ du_1$$

B is still an estimable variable of the sum; it is also an estimable variable of any subdivision of the system, if possible.

2.3 The estimation of estimable variables[8].

Consider now a system of known energy u. Let the prior density p(u/B) be positive for all positive u, as is usual. Since the energy is known, the very definition of estimable variables, as parameters, fails. In fact, any single system can be considered as being in "thermal equilibrium" with any thermostat (i.e. an infinite set of other systems in thermal equilibrium with each other), since the addition of a single system, having a possible value for u, does not perturb the probability distribution of an infinite set.

From the viewpoint of prediction of future events happening to the system, the population from which it is drawn should have, intuitiveley, little importance (we shall in fact assert later, after Szilard, that it has strictly no influence, and this will be shown possible only in the case of the Maxwell Boltzmann prior distribution). Strictly, to know B is therefore a problem of "retrodiction". But one may also wish to predict a prior distribution, for the energy of other systems of the set from which the first was drawn, and, for that, some reasonable guessing of the initial B would be useful. This B, although unknown, is not a random variable. Except under very rare conditions, where one has a prior distribution for it (the "Bayes case"), there is no limit-of-frequency sense, to be attributed to the feeling that "some values of B are more 'probable' than others". However, there is a very clear meaning to statements of probability of making an error by guessing a certain value for B. One can then try to construct estimators, or estimating intervals, so as to minimize certain chosen probabilities of error. Mathematical statistics is a technique for doing so, but it cannot ever justify any chosen criterion of good guessing.

Any estimator is a single-valued function of u; therefore, by applying it to parts of a whole of uniform temperature, one obtains a classical probability distribution of estimated "local temperatures": the parameter of this distribution is the true B : therefore estimates of B are not intensive variables. Many estimators are consistent, i.e. the distribution of the local temperatures gets more and more concentrated around the true value, when the size of the samples used increases. This makes it possible to measure the temperature of an infinite set with any precision. But for any sample size, the estimation of B is an essentially irreversible procedure; this fact was emphasized in the information theory literature by P.M. Woodward[12] (only in the Bayes case); it will be taken up in Part II in great detail.

Under these conditions, the replacement of the knowledge of u, and of p(u/B), by a single estimator of B, or by upper and lower bounds for B, is an operation entirely different from the measure of the random variable U. In the Bayes case, however, it is very close to the operation of prevision of the most probable or average future evolution of a system ruled by probability laws. The predicted evolution is of course different from the actual one. The point has lead to great discussion in quantum theory, see von Neumann[13] (§V.1.). But a simpler case of the difference between spontaneous random evolution, and noisy estimation is given in information theory by a comparison of Shannon's definitions of information of a Markovian message and of a noise-perturbed one. Both are specified by stochastic matrices (matrices such that line sums are one). But the ideas are entirely different, and the main point of reversibility and irreversibility in statistics is more basic than the hypotheses that matrices are stochastic (or unitary in quantum theory).

Fiducial distributions. As a measure of "degree of confidence" in various values of B, R.A. Fisher[14] (1934) has introduced a certain "conceivable" distribution of possible values of an actually non-random variable, a distribution based upon p(u/B). This concept (defined only in the case of sufficiency, see § 3.1) is often considered as to be carefully avoided, even if it is not claimed to be a kind of probability distribution. It seems however to be implicit in some thermodynamical arguments, where it is not necessary, as we shall see.

In view of this irrevesibility of estimation, we shall delay the consideration of specific methods to §5, and first consider the case where there exist intermediate steps of the estimation, which are reversible with respect to certain questions, that is : such that if one takes them, one loses "no information", in appropriate senses of information (more general than Shannon's). One definition of reversibility will fully determine the Maxwell-Boltzmann distribution, that is : characterize thermodynamics. But in any case, if there exist possible reversible steps, one should start by them.

3. The reversible part of the estimation of B. Derivations of the Maxwell Boltzmann distribution.

3.1 First derivation, based upon the concept of sufficiency.

Criterion of sufficiency. Suppose that the extensive observable u can be chosen to be additive : this is indeed a very strong hypothesis, since it excludes any interactions between neighbouring systems, and a fortiori quantum interactions between distant systems. Let us take several sample systems from a population believed of uniform B. Let u_i be their energies, and $u = \sum_i u_i$. One could estimate B, through B_e, either starting from u or from all the u_i. Independence of the result with respect to shape or disposition of the systems, implies that B_e should be a symmetric function of the u_i. If moreover any $p_i(u_i/B)$ can be considered as the distribution of a sum of still smaller variables (the possibility of doing so will be investigated in (Part III) one can estimate B_e from finer and finer subdivisions of the system. But B is considered as a "macroscopic" intensive variable, which roughly implied that from a certain level down, no improvement of the estimation can be obtained by measuring finer subdivisions, although such subdivisions are quite possible.

Postulate this independence of the estimate, relative to the knowledge of individual u_i, to be strict, and no more asymptotic. This means that, for example :

(S) $\Pr(u_1/u, B) = \Pr(u_1/u)$

independently of B. Nothing could then be drawn from the observation of the sample distribution of the energy among the different parts : no Maxwell's Demon could beat the macroscopic observer in the estimation of B, by measuring the partition of energy among molecules. In R.A. Fisher's[14] statistical terminology, u is then said to be a sufficient statistic for the estimation of the original B. If sufficiency is postulated at one level of subdivision, and additivity at all levels, sufficiency holds at all levels.

The intuitive concept of macroscopic was apparently weaker: it meant only asymptotic sufficiency : estimation depending rapidly less and less on finer subdivision; it could then depend on the first subdivisions. But on the other side, putting together several systems, and letting them interchange energy, is considered to be a reversible operation at any level. In fact, however, any axiomatic approach to thermodynamics is bound to be scale-less (see the axiomatics of Caratheodory, which is valid even for a single molecule[15]; see also the principle of Casimir, that the phenomenological laws of irreversibility apply even to fluctuations). It will turn out that strict sufficiency is implicit in the usual thermodynamics.

To sum up, the principle of sufficiency can be considered as established by all the experiments leading to the belief in the possibility of macroscopic description: It is thus established only below a certain level. On the contrary, the principle of additivity is established only above a certain level. We postulate that both hold exactly in a certain strip of sizes.

Maxwell Boltzmann (M.B.) distribution. Under certain regularity conditions, the only distribution for which u is a sufficient statistic for B is the M.B. probability density :

$$p(u/B) = G^{-1}(B)\, S(u) \exp(-Bu)$$

The detailed statistics can be thus derived from a single overall principle, a qualitative one, in the sense that it answers to a yes or no alternative: this cannot fail to recall Carnot's principle or the axiomatics of relativity theory. Strictly speaking, one should write some function of B instead of B, but the scale of B can be assumed fixed in consequence (the scale of u is fixed by additivity). One can even prove from sufficiency that the best estimate of any function of B is that function of the best estimate of B ; in the case of sufficiency, estimation and functional transformation are commutable.

The above result was proved independently in 1936, by G. Darmois (see appendix 1), B. O. Koopman[17] and E.J.G. Pitman[18]; before them, it was implicitly proved by L.Szilard[8] in 1925. Under slightly weaker conditions of regularity, the distribution need not have a density, and may be of the form :

$$dP(u/B) = G^{-1}(B)\, dV(u) \exp(-Bu)$$

S(u) and V(u) are the "structure function" and the "integral structure function" of the system. The "structure generating function" G is, because of the requirement that $\int p\,du = 1$, the Laplace transform

$$G(B) = \int S(u) \exp(-Bu)\, du = \int \exp(-Bu)\, dV(u)$$

we shall also need the function $J(B) = \log G(B)$, called "Planck potential". The usual probabilistic characteristic function of the M.B. distribution is

$$\varphi(t) = \int_0^\infty e^{itu} p(u/B)\, du = G(B+it)/G(B)$$

(The usual generating function is $\varphi(-i \log x)$)

Essentially, sufficiency, through the M.B. distribution, requires the separability of the variables u and B, appearing in the two-variable function p(u/B), in the sense in which one separates variables in differential equation.

The general result, that the expected value (E) and the variance (D) of U, knowing B, are the first two derivatives of $\log \varphi(t)$, for t = 0, now becomes :

$$EU = E(U/B) = \int_{-\infty}^{+\infty} u\, p(u/B)\, du = -\frac{d \log G(B)}{dB} = -\frac{dJ}{dB}$$

$$DU = D(U/B) = \int_{-\infty}^{\infty} (u - EU)^2 p(u/B)\, du = \frac{d^2 \log G(B)}{dB^2} = \frac{d^2 J}{dB^2}$$

Note that G(B) may be defined only for $B \geq B_d > 0$, and be a divergent integral for $B < B_d$. It is impossible in ordinary thermodynamics, of matter, that $B_d > 0$; but, clearly, the present theory is much more general than is required by matter, and there are exceptional applications where a positive abscissa of convergence of the Laplace transform is the main feature; see [1,2,3,4,5]

When J(B)/B is identified to the "free energy" of a system[15], it is seen that the free energy cannot be any function of B : its exponential must have a positive inverse Laplace transform (it is also called "completely monotone"; the signs of successive derivatives alternate). This fact has, surprisingly, never been mentioned, to our knowledge, in papers on statistical thermodynamics : it is clearly because for very large systems, it is irrelevant, cf. §6, so that there is no need to justify it in large scale thermodynamics.

Distribution of a sum of MB systems. Consider the sum of two systems following MB distributions with respectively structure and generating functions S_1, S_2, G_1, G_2 and same B. The distribution of the energy of the sum of these systems will be

$$p(u/B) = \int_0^u S_1(u_1) G_1^{-1}(B) \exp(-Bu_1) S_2(u-u_1) G_2^{-1}(B) \exp(-B(u-u_1)) \, du_1$$

$$= \left[\int_0^u S_1(u_1) S_2(u-u_1) \, du_1 \right] \left[G_1(B) G_2(B) \right] \exp(-Bu)$$

Thus, the distribution of the sum of MB systems of same B is still MB : in the addition, structure functions transform by convolution, generating functions simply multiply, and the J(B) functions simply add. This latter function may be considered as a particularly intrinsic measure of "contents". The same results clearly hold for finite sums of systems, and also for denumerable sums.

3.2. Szilard's derivation of the M.B.distribution

It turns out that the above derivation of the M.B. distribution could have been an inverted paraphrase of a derivation due to Szilard[8]. However, Szilard did not go much beyond this derivation, and his presentation could not readily profit from subsequent developments of mathematical statistics. Szilard's paper is often quoted, see von Neumann[13], but seldom analysed. Similar considerations can be found in a paper by G.N. Lewis[19], quoted by Fowler and Guggenheim[20]. Both authors wished to show how fluctuation phenomena can be introduced into classical thermodynamics, without destroying its structures and spirit. "The second principle loses nothing in rigour because of fluctuations, and in no way becomes an approximate principle : it melts into a higher harmony containing the laws of fluctuation."

Szilard considers systems in random evolution, i.e. such, that their properties at one time determine only probabilities at a later time; in particular, energy exchanges between systems in contact are only random. He assumes that the probability distribution of a system at an instant of time depends on one parameter only; the temperature, that systems which have long been in "thermal" contact and have exchanged energy have equal temperatures and may be considered to be in thermal equilibrium; finally, that systems in thermal equilibrium at one time remain in thermal equilibrium. The existence of a single number T, characterizing thermal equilibrium, is the "Zeroth principle" of thermodynamics, following Fowler (see 21) (in our more purely probabilistic model, we did not necessarily need to introduce energy exchanges) (The fact that energy does not vary in the interactions, is the First Principle). Temperature is not however assumed to determine a single energy, but only some superposition of states of different energy (these states have an objective existence, independently of any observer); this distribution is realized among systems having long been in long thermal contact with a "thermostat".

The infinity of possible energy distributions will now be reduced, by translating the fact that p(u/T) cannot be modified without a compensation, for which a new thermal contact is necessary. This will be translated into postulates, about the more detailed nature of thermal equilibrium, which are a priori only "reasonable" but a posteriori experimentally correct, in the sense that they lead to the M.B. distribution.

Let two systems, of energies u_1 and u_2, initially in contact with a thermostat, be separated and brought in very long thermal contact with each other. Their initial energies U_1 and U_2 were independent random variables, of same parameter T. Szilard assumes that the nature of thermal equilibrium is such, that the energies U'_1 and U'_2, after long contact, are again independent random variables, having moreover distributions independent of the initial temperature, and of initial values u_1 and u_2, and conditioned only by the constant total energy u.

Similarly, Lewis considers that, when a quantity is shared between two systems, "the ratio of specific probabilities of any couple of partitions depends only upon the nature of the two systems, and in no way upon their method of connection, or upon the existence, nature or mode of connection of other systems." He then postulates that "one should expect that the probabilities of various partitions of a quantity U between two systems depend only upon the total u, and in no way upon the reservoir with which the two parts are in connection. Thus, one would expect the same partition, whether the two systems are in very imperfect contact with one reservoir, or whether they are in contact with another reservoir, of very much higher temperature, but have the same total energy through a very rare fluctuation. (This is not an exact quotation, but the English translation of the author's French translation of Lewis's words).

It is seen, with great pleasure, that both authors have rediscovered the principle of sufficiency, exactly in the form (S) of § 31, but have interpreted it as a property of equilibrium, not of observation. As already mentioned, Szilard has even anticipated the derivation of the M.B. formula, and Lewis has considered ratios of probabilities of couples of partitions, which is now a constant procedure in likelihood ratio tests.

It is remarkable that one can use such a "yes or no", and not extremal definition of equilibrium, and go so far.

To prove the MB pdf, Szilard writes that $p_1(u_1)\, p_2(u_2)/p(u_1 + u_2)$ should be independent of the T. Compare this with some "derivations" of Boltzmann's other formula for entropy S: "S = log p". They amount to assuming that $p_1(\varepsilon_1)\, p_2(\varepsilon_2)/p(\varepsilon_1+\varepsilon_2)$ is not only independent of the T, but strictly equal to 1. This has the disadvantage of a poorly defined probability, and of the immediate introduction of a new concept, identified only later. Besides, to introduce the needed S(u) term, one must postulate immediately that one energy may correspond to several different distinct (degenerate) states (see § 5.1) Szilard's approach is clearly preferable.

Through a discussion of fluctuation sizes for different systems, Szilard shows that T is a universal temperature; however his discussion of fluctuation is less complete than it is now possible to make, and generally, his approach, although closer to usual thermodynamical thinking, is less easy to generalize (chiefly because of the restriction on intensive variables).

Principle of Onsager-Casimir. In the theory of the irreversible decay of deviations from equilibrium[22], one has to assume that the same laws apply to "small but macroscopic" deviations, and to fluctuations. This amounts to postulating that the way a deviation was reached is irrelevant, and only its amplitude imports. This is a markovian hypothesis[23] (which is a probabilistic form of Huygens' principle): statistical independence of past and future, when the present is known. It leads to Onsager's relations of reciprocity. In the necessarily weaker theory of equilibrium, we are concerned with, one makes a weaker hypothesis that independence from the past is attained after a long time only. Thus, Onsager's theory is a quite proper "interpolation" of the present theory, with a full markovian hypothesis. It is curious to note that a weakened probabilistic Huygens' principle leads to the same results as Carnot's principle.

3.3. Second derivation, based upon the concept of efficiency.

Criterion of efficiency. It was mentioned in §3.1, that, since the estimate of B is a non-random function of u, it is a random variable, when referred to the ensemble of constant B, from which the system is drawn. Assume that the distribution of the u is known, but not necessarily M.B., and compare different possible methods of estimation, all assumed to be unbiased, that is such that the mean of the estimator is equal to the true value. For that, compare their variances, i.e. mean square deviations from the mean. (This is a quite different concept from the usual temperature fluctuation).

It can be shown (see appendix 2) by an application of Schwartz's inequality, that, whichever the distribution p(u/B), the variance of an estimate of a function f(B) is necessarily bounded below by the expression :

$$D[f(B)] = \frac{\left(\frac{df}{dB}\right)^2}{F}, \text{ where } F = E\left(\frac{d \log p(u/B)}{dB}\right)^2 = E\left(\frac{d^2 \log p(u/B)}{dB^2}\right)$$

F is called "Fisher's information". This limit means not only that one does not know how to perform a more precise estimation, but that under certain conditions of regularity, one could not conceive of any such estimation (which, even though it could not be justified by its properties, should not be dismissed for empirical reasons). (Remark that Hodges and Le-Cam have shown examples of estimators more precise than the above limit, but only for a set of values of B which is of Lebesque measure O). The existence of upper limits to the precision of measurements considered as indirect, which are in fact estimations, is quite familiar in quantum theory.

Maxwell-Boltzmann's distribution. Usually, there are still closer lower limits to variance, but Fisher's limit can be exceptionally attained (Appendix 3) if

$$\frac{d^2}{dt^2} \log E\left(e^{itu}\right)\Big|_{t=0} = E\left(\frac{d^2 \log p(u/B)}{dB^2}\right)$$

Under certain conditions of regularity, the lower limit to variance can be attained only with the M.B. distribution. Thus the overall statistics can also be derived from a variational principle.

Then, Fisher's information takes the especially simple form

$$F = \frac{d^2}{dB^2} \log G(B) = \frac{d^2 J(B)}{dB^2} = D(E)$$

it is a function of the "energy content" J(B).

Uncertainty relationship. Let f(B) = B. Then clearly

$$\boxed{DU.\ DB = 1}$$

The less well-known is the energy of a system, when its true temperature is known, the "larger" the system (in terms of its contents log (B), and the better the temperature can be estimated from such a large sample. And conversely.

Or else, disregard the true temperature. Any estimation gives B together with DB; identify the estimator of B with the true B, but only for the purpose of estimating the DU of U before the measurement. Then, the larger DU, the smaller DB.

One cannot fail to note the formal identity of this relationship with Heisenberg's relationship, or its equivalent in communication: Gabor's relation "Df. Dt = 1/2" in the spectral analysis of signals. In fact, all three are one relation : they result from equality cases of Schwartz's inequality, for dual variables. It does not matter, of course, whether the duality is Fourier, (quantum theory and spectral analysis) or Laplace (here). But in fact there is a deep difference : in Heisenberg's relationship, the observer can choose which variable he wants to know with greater precision; here, this is determined by the contents J(B) of the system studied : B is known exactly only for the infinite thermostat, and U exactly only when it is certainly zero.

However, it is pleasing, writing DU.D(1/T)=k, to find an absolute meaning attached to size, in a way formally quite similar to the introduction of the size of quanta by the relation DU.Dt = k.

Proper scale. It would be pleasing to have a new scale of B, such that DB be independent of the estimated B, that is, from the measured u. (This seems to be what MacKay calls the "proper scale"). Clearly :

$$f_p(B) = \int^B \sqrt{\frac{d^2 \log G(B)}{dB^2}}\, dB$$

This proper scale was implicitly used by Girshick and Savage[24], in proving that u is an admissible minimax estimate of the intensive variable EU, relative to the risk function $(u - EU)^2/DU$.

"Noise". Take now $f(B) = EU = -d \log G(B)/dB$ One finds that D(EU) = DU. In other terms, the estimation variance of the "true" value EU, is equal to the fluctation of u around its mean value. This is not obvious, but a theorem, which should be brought together with the fact, mentioned above, that when a sufficient statistic exists, the estimate of a function of B is the function of the estimate. Further, the fact, that even when u is assumed strictly known, there is a sense to be attributed to the fluctuation of something very close to U, is very comforting, in view of the paradox in considering the measure as infinitely precise. But even this new noise has nothing to do with "hidden variables".

3.4 Other derivations of the MB distribution

The concept of sufficiency has many other aspects. Their discussion has better however be delayed until some new concepts have been introduced. Let us mention however, that sufficient valuations of observations preserve "information" in many different senses of the word, Fisher's, and also Shannon's (and in a generalisation of both these senses due to Schutzenberger, see Part II), and also in a sense due to Bohenblust, Shapley and Serman (see §5.2).

4. Discussion on the methodology of §3 : Two complementary types of statistical thermodynamics.

Let us interupt here the construction of the theory, to comment upon what is being done. (The reader may immediately proceed to §5, where a great amount of new arbitrariness is introduced, in the actual irreversible estimation of the B of the M.B. distribution, and to §6, where this arbitrariness is shown to asymptotically vanish for all practical purposes, in large systems.)

We have succeeded (partly after Szilard) in deriving, from purely phenomenological criteria, some results of statistical thermodynamics, a science considered to be so deeply related to kinetic and such models, as to be also called statistical mechanics. The M.B. distribution is no longer characterized as being the one which would be steady under collision phenomena, but as having certain good properties under observation. Several non classical results were also found. How does this fit into the classical scheme?

4.1 The two classical methods of thermodynamics.

At least since Clausius, one recognizes two methods of mathematical structuration for thermodynamics: the phenomenological, also called macroscopic, pure, classical, axiomatic, etc., and the kinetic, also called microscopic, statistical, etc... The latter contains more results than the former, in particular the statistics, and is also the older of the two (the Greeks, Gassendi, the Bernoullis), and closer to intuition. Despite this, it is considered as conceptually subordinate to the other, since it is required to derive the principles of the other as theorems, whereas the reciprocal is usually not considered.

How well is this explanation achieved? From the viewpoint of rigor, notoriously very poorly: there is in fact a complete paradox a priori in any kinetic explanation: it is the fundamental incompatibility of structure between the irreversibility of some principles of phenomenological thermodynamics, and the mechanical reversibility of any purely kinetic model, one could ever think of for these principles, (e.g. the paradoxes of Loschmidt and of Zermelo[25]). These objections gradually forced Boltzmann to add to the mechanical assumptions. Following Uhlenbeck, let us call the problem of reconciling the two viewpoints, Boltzmann's problem[26]. The ideal would have been to derive the macroscopic results from microscopic assumptions; but in fact one[27] is more properly looking for analogies, or logical structures analogous to thermodynamics, by adding to the kinetic models some rather arbitrary randomness assumptions, often introduced through the necessary inprecision of

"coarse observation": Any initial unevenness of probability distribution, although strictly preserved, because of Liouville's theorem, is supposed to dissipate itself into thinner and thinner streamlines; so that any coarse density becomes uniform, after an average time which increases when the coarseness decreases. For a system of a given age, there will be an extremely rapid, and increasingly sudden, variation of properties, when the size of the cells goes to zero. For a very old system, one tends towards the situation, in which there is a sharp discontinuity, and a "singularity" of phenomena near infinitely sharp definition. This recalls the singular phenomena observed when the viscosity of a fluid tends to zero, which also relate to irreversibility and dissipation. Once established, this randomness can be followed in its development, with no new conceptual paradox, although with great technical difficulty, there is a great amount of current work on this topic[28]. But the introduction of randomness still raises entirely uncleared problems.

4.2 Statistical thermodynamics without hidden variables.

Since, therefore, the kinetic foundations of thermodynamics are not sufficient, without a further (single[29]) hypothesis of randomness, are they still necessary in the presence of such a hypothesis? After all, although statistical thermodynamics owes its origin and development to the theory of atomism, it need not always be so. In fact, partly after Szilard[8], we are in the process of showing that one can "short circuit" the atoms by centering upon any element of randomness, for example introduced through necessarily unprecise observation, and we are deriving a substantial part of the fundamental results, usually obtained through kinetic arguments, by following up the introduction of randomness in a purely phenomenological way.

The fact would have been of course more striking, before the times when atomic theory ceased to be a rather doubtful conjecture. It is also a pity, that Szilard's original paper was not more often quoted in contemporary (around 1925) discussions about the "causal interpretation" of quantum theory through "hidden variables". In fact, a rapid decline of interest in the phenomenological and "energetist" approach to thermodynamics was contemporary of the Copenhagen approach to quantum theory(the greatest success of the school of "never going beyond observation", even in words: a still triumphant school in thermodynamics in 1895, when Boltzmann was complaining that "kinetic theory was, so to speak, out of fashion in Germany"). There is however a renewed interest now in causal reinterpretations. Take for example von Neumann's[13] proof of the impossibility of introducing hidden variables, into the conventional (Copenhagen) approach to quantum theory, without first modifying it. This proof is part of a reduction of a large part of the quantum rules to a set of phenomenological and axiomatic rules for observation, starting from a purely stochastic viewpoint very similar to that used here. The incompatibility between a quantum structure and hidden variables, may therefore be compared to the incompatibility between thermodynamics and kinetic theory (paradoxes raised against Boltzmann). An attempt to go around those paradoxes, such as current work of Bohm[30], de Broglie[31] and Vigier[32], is to be compared to the attempts to "solve" Boltzmann's problem. This current work does not actually attempt to build a theory of which the quantum theory would be the "thermodynamics", but simply to disprove the impossibility of building such a theory, by introducing randomness not fundamentally, but through a chaos hypothesis. It may be that the possibility of such a twin set of quantum thoeries would now appear less shocking in view of the existence of a statistical thermodynamics without hidden variables, apparently just as "closed" as the conventional quantum theory. However, even if there some pleasure and help, in the great similarity of the words used to describe the two situations, there is no question of mathematical identity: an opposition will remain between the unitary and contact transformation, and between a purely real and a complex theory. Only the single opposition, of quantum vs. non quantum statistics, will be completed by a second opposition, of phenomenological vs. kinetic. Each of the six possible comparisons of the four theories is enlightening.

This methodological discussion will be continued through §6 and in the conclusion to part I

5. The irreversible and final part of the estimation of B.

5.1 First method of estimation: maximum likelihood B and Boltzmann's most likely state. Degeneracy.

To derive the M.B. distribution, in §3, we did not need to specify the actual arbitrary method of estimation, to be used in order to attain the optima shown to be possible. The function linking B to u must now be specified. In principle, rules of estimation should be derived from the properties desired from the estimate. In fact, however, the most usual rules of statistics were introduced for no conscious reasons, except simplicity, and if motivated at all, were first motivated by considerations of "degree of reasonable belief" quite foreign to classical probability; and only later justified by their properties[7]. The estimation theory of thermostatistics is only implicit, though quite real; it will turn out, curiously enough, that it has used exactly the same procedures as statistics!

Start by the most widespread theory of estimation: R.A. Fisher's[14] theory of maximum likelihood estimation. To find a measure of rational belief in a value of B, when we are reasoning from the sample to the population, Fisher inverts the function p(u/B), taking now u as a parameter, and B as a variable. Of course, p(u/B) ceases then to be a probability distribution, and B is no random variable. For example $\int p\,dB \neq 1$ (and it even depends upon the arbitrary scale of B); therefore one cannot speak of the likelihood of the set of all values of B, but only compare likelihoods. Taking now the M.B. distribution as having been derived in §3, maximize

$$\log p(u/B) = -\log G(B) + \log S(u) - Bu$$

This requires that

$$u = -\,d \log G(B)/dB$$

B will be obtained by inverting this implicit equation.

One cannot fail to note the identity of this result with a classical formula by Boltzmann[15]. Let us review the proof of that result. To define the temperature of a system of given u, one takes that only EU = u is known, then one derives the "most likely" distribution of this energy between the available (discrete) states. For that, one maximizes the "entropy", or logarithm of the likelihood, given EU = u, and El = 1. The derivation, using Lagrange multipliers, is too classical to repeat. The reason for saying "likelihood" instead of probability, will appear soon. This gives the distribution

$$p(\text{any state of energy } u/B) = Ke^{-Bu}$$

as being the most likely (M.L.) This seems to differ from the M.B. distribution, but in fact one introduces the further assumption of "degeneracy": that there may be $S(u)$ different states of same energy. Then ,the most likely distribution of energy turns out to be the MB pdf; without further arbitrariness, one gets a relation between B and EU, viz :

$$EU = -\,d \log G(B)/dB$$

Then one replaces herein E(U) by the known u, and inverts to get B. Altogether Boltzmann's argument amounts to an implicit and improvised theory of estimation, anticipating Fisher's theory. Note that the assumption of degeneracy would not have changed the result of taking the M.L. value of B, in Fisher's approach, since the probability of any single state of energy u is

$$p(\text{any state of energy } u/B) = G^{-1}(B) \exp(-Bu),$$

and since the log $S(u)$ term drops out in the maximation of log p, relative to B.

Therefore, taking degeneracy for granted, the only technical progress made, comes from the fact that maximum likelihood estimates have been rather more carefully studied than most likely states. It is quite true that the M.B. distribution is now obtained without the arbitrariness of the "maximum" specification (which needs no amplification in thermostatistics, see Darwin and Fowler and Khinchin[33]; but the same arbitrariness is found immediately at the next step.

But what about a conceptual benefit taken from the translation? Neyman[7] had critically shown, in detail, in what respects a likelihood differs from a classical probability as a limit of frequency; the same defect must be present somewhere in Boltzmann's likelihood, called a probability and distrusted as such, but never analysed as carefully. Of course the expression

$$Q = \prod_s p_s^{f_s}$$

(where f_s is the frequency of the state s, in a finite population of N systems) is a probability, that can be multiplied for independent events, and added for disjoint events, and can be considered as the product of probabilities of the individual events that each of the N systems be in the state it is in. True also, Shannon's fundamental lemma of information theory (a corollary of the weak law of large numbers), asserts that, for large enough N, and except for a set of events of total probability as small as one wishes, the probability of all distributions of energy is as close as one wishes to

$$\exp(-NH) = \prod_s p_s^{Np_s}$$

The number of "complexions" for which there are f_s systems in state s, being $C = \prod f_s!/N!$, the total probability of complexions characterized by f_s is QC. It is seen to attain a maximum if $f_s = N p_s$. But the probability of the most probable partition is not QC, but $(\max p_s)^N$. Besides, there is no sense in considering as a product of N times the "probability" $\prod p_s^{p_s}$; the latter expression, not involving asymptotics, is not a probability of any event. At best, it could be the "probability" of a probability distribution, but formed in its own probability field! But mostly, it ceases altogether to be a probability, when one compares different sets of p_s, and becomes a likelihood, in the exact Fisher's sense. In fact, this likelihood seems to be even more general than in the above application, since it is not limited to a parametric family of distributions; but a posteriori, Lagrange multipliers show that the most likely distribution is bound anyway to be one with a single parameter. Note that the comparison of probabilities of two distributions in their own fields is made less absurd by that: $d\sum p_n \log p_n = \sum dp_n \log p_n$: actually one chooses one of the two fields, as terrain of comparison. Note also another dubious feature of the probability arguments involved in Boltzmann type of maximation, which appears in quantum theory: one does not obtain probabilities, but probable values of the numbers of appearances of each alternative.

A final defect of maximation derivations of the MB pdf, by contrast to the sufficiency derivation, is that successive applications of Stirling formula and passages to the limit hide the exact domain of applicability of the result obtained.

5.2 Second method of estimation of B : the Bayes method. Choice of S(u), Minimum risk.

Bayes estimates. The oldest method of estimation, against which the maximum liklihood method was initially set up, was Bayes' s method, based upon the assumption that,(on the basis of reasons of believing that some values of a parameter are more"frequent"than others) one may set up prior probabilities for B, which thus becomes a random variable. Then one may compute posteriori probabilities, and estimate B to be the mean or a posteriori most probable value. Note that one may multiply the prior probabilities by any constant, without changing the posterior mean or most probable value. Sometimes this is used to justify using "weights" for the different prior states, which do not add to any finite number: the axiomatic basis for such conditional probabilities was given by Rényi.[34] Recall that Boltzmann' s method did not involve any prior probabilities: it is another reason for considering an improper its specification as a theory of the "most probable".

The belief in the necessity of prior probabilities in estimation(in "going from effects to causes") is very widespread in physical papers. This was von Neumann' s point of departure in his 1932 axiomatization of the quantum theory; he also notes that, since there are different possible prior probabilities, the problem does not have a unique solution; but finds that a certain prior distribution is "distinctly different from others", and bases the whole theory on it. Gibbs on the contrary warned of the impossibility of determining probabilities for prior events on the basis of posterior ones. Others, such as Fowler, use widely the formalism of weights of the Bayes theory, but insist that these have nothing to do with the time a system spends in a state.

In general, Bayes estimation is further based upon uniform prior probabilities; this is supposed to result from a principle of "unsufficient reason". There seems therefore to be a trace of the Bayes appraoch to physics any time that cells are assumed equiprobable. Consider as an example the method of estimation of B, which is implicit in Darwin and Fowler' s method of mean values. There, one takes the mean value of, say, energy, assuming all partitions of energy to be equiprobable. Then, one finds that a parameter B introduces itself through a mathematical trick, and altogether one estimates this parameter as being the one for which the mean value of energy were the observed value.

Choice of S(u). There is no need to give here a list of modern methods of statistics, free from both prior probabilities and likelihoods, since we are rather reviewing the conventional thermostatistics, in the light of the kind of statistics it involves. One should note however that Bayes' thinking, in the form used by von Neumann, is also present whenever one justifies the MB distribution, as being invariant by shocks of a certain class . This is so, because, from the view-point of the kinetic theory, the energy is no random variable, as we have stressed it in §4, and making it such is as criticable, as making a parameter into a random variable: to stress the analogy, imagine that the "probability a priori" is that of a distribution of energy at an instant of time, while the "probabilities proper" are the transition probabilities between different distributions. Therefore, whichever the progress of statistics, the critique of Bayes' s approach, in a given time, will be required in thermostatistics. Since our approach is not kinetic, this is not a problem here, but in fact Bayes thinking may also be used to determine (or rather "choose") the structure function S(u). The choice of this function by the observer, who knows u, is a basic aspect of the arbitrariness of estimation of B, which we have not stressed enough so far, since in principle it does not belong to estimation, but to the construction of physical models. The present point is that such models are often based upon "unsufficient reason" arguments, akin to those used in Bayes estimation. In fact, the whole theory of degeneracy is based upon it, since to identify the structure function S(u), with a number of states of given energy, (a number that can be calculated from geometric and combinatorial considerations), one may postulate that when the energy available, and therfore B, tend to infinity, the ratio of the probabilities of any two states must tend to one. This is close to "unsufficient reason", the list of states changes when the "reasons" change. (quantum statistics).

A remark on sufficiency and smallest risk

Consider now all those criteria of estimation, which aim to minimize a certain average(Bayes) "risk", linked with occasional wrong estimates. Under those conditions Bohnenblust, Shapley and Sherman have defined an experiment "a" as being not less informative as experiment "b," if any risk attainable with experiment b, could also be attained with experiment a. Blackwell and Girshick[35] have used this criterion to compare the measurement of the U_i of a set of parts, and of the total U alone, when u is a sufficient statistic. He has shown that if the number of alternatives is finite, the measurement of u is not less informative than that of the u_i' s. The

result is most presumably still true for a continuous infinity of alternatives, under some regularity conditions, which may be the same as those leading to the M.B. canonical distribution.

E. Lehmann has considered tests of hypotheses, such that "B is in a certain interval". He shows that when a sufficient statistic exists, any test based on a function of the measures, may be replaced by a test based on the sufficient statistic, without increasing the power function, that is, the probability of wrongly accepting the hypothesis.

Also mention the following phenomenon (for definitions and proof, see Kendall[30] II 281): "If a system of uniformly most powerful tests exists, and if any point in the sample space lies on the boundary of a best critical region then a sufficient estimator exists for the parameter whose variation provides the admissible alternatives)".

§6. The arbitrariness in estimation vanishes asymptotically.

6.1. Passage to the limit of very large systems, i.e. very large log G(B).

Since there is no motivation a priori, in classical probability theory, for the Boltzmann - Fisher estimates of "rational Belief", these estimates should be judged by the properties they turn out to have. In fact, they have extremely satisfactory asymptotic properties.

In thermostatistics, they converge to the Darwin Fowler estimates; when the size of the system tends to infinity. This was proved by Darwin and Fowler by the method of steepest descents; later by Khinchin, by the method of local limit theorems of probability. Therefore, since the physicist is not interested in small bodies, the Boltzmann procedure is justified by its simplicity, and by its convergence to theories considered as "quite correct."

In statistics, they behave well for large samples. Chiefly, they are consistent (they converge to the true value, with probability approaching unity, as the size tends to infinity); they are normally distributed around the true value, and their variance is the smallest possible (efficiency). Given the fundamental arbitrariness of estimation criteria, one could not say that any other procedure is more correct than this one; so, whenever it is the simplest, (for example when a sufficient statistic exists) and since it is at least as good as any other, one can use this procdeure for large samples.

6.2. Small systems

As their main drawback, from the viewpoint of present day problems of statistics, maximum likelihood estimates lack reasonable small sample properties; and among the class of consistent, asymptotically normal, efficient estimates, they may not be the ones that converge fastest. The statistician cannot any more justify the definition of an estimation procedure by its asymptotic properties.

However, the physicist is not troubled by this, and continues to take the best from the facilities, permitted by the "fortunate insensibility of thermodynamic functions", stressed by Lorentz.

6.3 Possibility of constructing observation models of thermostatistics.

This fortunate insensibility is the very key to the possibility of constructing models in physics, which are based upon conditions of optimality of observation, despite the logical arbitrariness of what is "the most favourable". In the general case, this is impossible: the criterion of an engineering problem can be changed at will, without its ceasing to be another acceptable engineering problem. But the state of nature, most favourable for one problem, may not be so for any other, so that the fact, of being most favourable, is not in general a way of describing a real physical situation. One may say that each time one describes (and can fully characterize) outside circumstances, as being the most favourable for somebody, one builds a new anthropocentered physics; (this is also the case when one assumes that the opponent in a zero sum two person game plays a minimax strategy against you, choosing your matrix of imputations) To invert methodologically an engineering problem, is to identify this physics with an actual one. But how to defend the model so obtained, if the criterion of the engineering problem is irreducibly arbitrary?

The whole point is that arbitrariness exists anyway in thermostatistics, and that is exactly the same than in statistics. This is in no way surprising, in view of the way in which probabilities were asisgned to hypothetical ensembles in thermostatistics; to "agree with our partial knowledge" (a fact linking irreversibility not to incompleteness or inexactness of mechanics, but to incompleteness of specification). Criteria of equipartition are only "reasonable" criteria, never considered as anything but mathematical abstractions of reality, or in other words as "fascinating" mathematical tricks, but still tricks; and they may become very dangerous conceptually when "asymptotically slight" differences go undetected altogether, and entirely different concepts get exchanged

because they lead to the same numerical results. Entropy is the worst sinner in that respect, and it will be studied in that light in §7 (second part of this paper).

It is a pity, of course, that the phenomenological and statistical approach can only be franker about doubtful points, can help circumscribe and better understand them, but cannot help eliminate them. Statistics may be sorry for it too, because it makes hopeless any belief in an improvement of the choice among statistical criteria, on the basis of their appropriateness for physics.

Conclusion to part I

In part II of the paper, we study the problem of the informations and entropies; in part III, the problem of the possibility of interpolative subdivision of physical systems, and in part IV, the generalization of the whole to systems with exchange of energy, to quantal systems, and to non-parametric systems. However, it can already be seen how thermostatistics has followed a development parallel to mathematical statistics. Recall how late (Khinchin[33]) the parallel development of probability theory and of termostatistics have converged; statistics and thermostatistics have been even slower, although statistics and probability have long converged, and could have perofrmed the link.

In the comparison, statistics appears as by far the more advanced discipline. Thus mathematical and methodological help turns out to go the "wrong" way, up the scale of sciences of Auguste Comte, since the modern small-sample statistics was made necessary in biology, agriculture, social science and industrial acceptance (where asymptotics are of no help, and the process of observation of so great importance that te there is no difficulty in accepting theories centered around the observer). Statistics became necessary for the physicists only when the problem of noise reduction had to be faced: this is the historical reason for the context of this reasarch.

It is in a sense very discouraging to have found on this occasion, that the wish for conceptual coherence of the theoretical physicist had been weaker than the practical needs of "automatisation of thought processes in inference" ("cybernetics"?); The moral is that the foundations of thermostatistics should be reviewed critically, after each important progress in the mathematics or methodology of other statistical disciplines.

Acknowledgements. The author wishes to thank the Rockefeller Foundation and the Centre National de la Recherche Scientifique (Paris), for research grants, during parts od the time while this paper was written; and Miss Knight, for secretarial help.

References of Part I

1 B. Mandelbrot : Simple games of startegy, occuring in communication through natural languages; Trans. I.R.E. Inf. Theory 3 (1954) 124

2 B. Mandelbrot : Statistical structure of language; Communication theory; Academic Press (1953) 486

3 B. Mandelbrot : Jeux de Communication; Publ. Inst. Stat. Paris 2 (1953) 3

4 B. Mandelbrot : Structure formelle des textes; Word (New York) 10 (1954) 1

5 B. Mandelbrot : Thermostatistics of Willis systems; Information Theory; Academic Press, (1956)

6 C.E. Shannon : Mathematical theory of communication; Bell S.T.J. 1948

7 J. Neyman : Lectures and conferences on mathematical statistics and probability Department of Agriculture (1952), chap. 4

8 Leo Szilard : Uber die ausdehnung der phenomenologischen thermodynamik auf der schwankungserscheinungen; Z. Physik 32 (1925) 753

9 J.L. Lawson, G.E. Uhlenbeck : Threshold signals; Mac Graw Hill (1950)

10 A. Kolmogoroff : Bull Ac Sc URSS Math 5 (1941) 3

11 N. Wiener : Times series; Wiley (1949)

12 P.M. Woodward : Theory of observation TRE Journal (1949)

13 J. von Neumann : Mathematica principles of quantum mechanics, Princeton (1954)

14 R.A. Fisher : Contributions to mathematical statistics, Wiley (1950)

15 A. Sommerfeld : Thermodynamics and statistical mechanics; Academic Press (1956)

16 G. Darmois : XXII Session de l'Inst. Int. Stat. Athènes (1936)

17 B.O. Koopman : Trans. Am. Math. Soc. 39 (1936) 399

18 E.J.G. Pitman : Proc. Camb. Philo. Soc. 32 (1936) 567

19 G.N. Lewis : J. Am Chem Soc. 53 (1931) 25

20 R. Fowler, E.A. Guggenheim : Statistical thermodynamics; Cambridge (1952)

21 E.A. Guggenheim : Thermodynamics; North Holland (1949)

22 de Groot : Thermodynamics of irreversible processes; North Holland

23 L. Onsager, S. Machlup : Phys. Rev. 91 (1953) 1505

24 M. A. Girshick, L. J. Savage; Second Berkeley Symposium Stat. (1951) 53

25 P. T. Ehrenfest; Encyclopedie Sc. Math. IV 1, 6 (1915) 188

26 G. E. Uhlenbeck; Lectures at Princeton, Les Houches, etc... (1955)

27 J. W. Gibbs : Statistical Mechanics : Yale (1903)

28 M. Kac : Third Berkeley Symposium Stat. III (1956)

29 L. van Hove : Physica 21 (1955) 517

30 D. Bohm : Phys. Rev. 85 (1952) 166

31 L. de Broglie : Comptes-rendus since 1952

32 J. L. Vigier : Interprétation causale de la mécanique quantique; Gauthier Villars (1956)

33 A. I. Khinchin : Statistical Mechanics; Dover (1949)

34 A. Rényi : Acta Math. Hungarica (1955)

35 D. Blackwell, M. A. Girshick : Theory of Games and Statistical Decisions, Wiley (1954)

36 M. Kendall : Theory of Statistics; Griffin (1948)

37 C. R. Rao : Advanced Statistical methods; Wiley (1952)

Appendices

A. 1. (after [37], § 4a. 5)

Let us assume that u_i are independent observations from a population, with range independent of B. It can be shown that a general necessary and sufficient condition, that a distribution admits a sufficient statistic is the Fisher-Neyman relation :

$$\Pi\, p\,(u_i/B) = P\,(Q(u_1, u_2, \ldots u_I), B).\ R(u_1, u_2, \ldots u_I)$$

where P (Q, B) is the density of the statistic Q, and R, is independent of B.

Assuming all functions partially differentiable with respect to B, write :

$$E\ \frac{d\log p(ui/B)}{dB} = \frac{d\log P(QB)}{dB} = W(QB)$$

Since this holds for all B, any value of B can be substituted, to obtain the relation

$$v = \sum v(u_i) = V\,(Q)$$

connecting Q and the statistic $v = \sum v(u_i)$. If T(B) and v(u) are differentiable functions, it follows that

$$\frac{dv}{du_i} = \frac{dv(ui)}{du_i} = \frac{dV(Q)}{dQ}\ \frac{dQ}{du_i}$$

Also,

$$\frac{dW(QB)}{dQ}\ \frac{dQ}{du_i} = \frac{d^2 \log\, p(ui/B)}{dB\,du_i}$$

Therefore, for all i,

$$\frac{d^2 \log p(uiB)}{dB\,dui} + \frac{dv(ui)}{dui} = \frac{dW(QB)}{dQ} + \frac{dV(Q)}{dQ} = L(B)$$

a function of B only; Integrating with respect to Q,

$$W(Q, B) = L(B)\ v(Q) + M(B)$$

and then with respect to B,

$$E \log p(ui/B) = X(B)V(Q(Ud, \ldots u_I)\ + Y(B) + A(u_I \ldots u_I)$$

$$p(u\ /B)\ = G'(B)S(V(u)\exp(-BX(B).v(u))$$

choose the scales so that X(B) be B, and v(u) be u : v is clearly additive.

A. 2 (after Rao [37], § 4a. 2)

Let $\Pi\, p\,(u_i)\, B)$ be the probability density of the observations and $V\,(u_1; \ldots u_I)$ be an unbiased estimate of f(B), a function of the parameter B of the density. Then :

$$\iiiint V(u_1 \ldots u_I)\ p(ui/B)\, du_1 \ldots du_I = f(B)$$

If the limits of integration do not involve B, differentiation under the integral yields :

$$\iiiint V(u_1, \ldots u_I)\ \frac{d\ p(ui/B)}{dB}\ du_1 \ldots du_I = \frac{df(B)}{dB}$$

By Schwartz's inequality :

$$\iiiint \left[V(u_1 \ldots u_I) - f(B)\right]^2 \Pi\, du_1 \ldots du_I \times \iiiint \left(\frac{d\log \Pi}{dB}\right)^2 \Pi\, du_1 \ldots du_I \geq \frac{df(B)}{dB}^2$$

Then taking I = 1, the Fisher's information inequality follows :

$$D(U) \geq (df(db)^2/F$$

A. 3 (after Rao [37], § 4a. 3)

The only equality case od Schwartz's inequality is when :

$$V(u_1 \ldots u_I)\ - f(B)\ = L(B)\ \frac{d\log \Pi}{dB}$$

where L is a function of B only. This is a differential equation, having the solution

$$E \log p(ui)B) = \int \left\{\frac{V(u_1 \ldots u_I)}{L(B)}\ - \frac{f(B)}{L(B)}\right\} dB = A(u_1 \ldots u_I) + VX(B) + Y(B)$$

where X(B) Y(B) are functions of B, and A is independent of B. Hence

$$p(u/B) = G^{-1}(B)S(V(u))\ \exp(-V(U)X(B))$$

It is easily seen that reciprocally, the lower limit to variance is attained by this distriction.

A RADAR DETECTION PHILOSOPHY

William McC. Siebert

Research Laboratory of Electronics and Lincoln Laboratory
Massachusetts Institute of Technology, Cambridge, Mass.

Abstract

This paper is an attempt to present a short, unified discussion of the radar detection, parameter estimation, and multiple-signal resolution problems--mostly from a philosophical rather than a detailed mathematical point of view. The purpose is, hopefully, to make it possible in at least some limited sense to reason back from appropriate measures of desired radar performance to specifications of the necessary values of the related radar parameters. Specifically four measures of performance quality are considered:

1. The reliability of detection,
2. The accuracy with which target parameters can be estimated,
3. The extent to which such estimates can be made without ambiguity,
4. The degree to which two or more different target echoes can be separated or resolved.

It is argued that the radar synthesis problem can be split into two more-or-less independent phases. First, adjust such parameters as those appearing in the radar equation so that the received signal energy is sufficiently large for the degree of reliability of detection desired. The required value of energy is almost entirely independent of the character of the received echo signal waveform. The second phase is, then, to select the waveform in such a way that accuracy, ambiguity, and resolution requirements are met. The limitations on what can be achieved in terms of these three quality measures are discussed in relation to an uncertainty principle. For purposes of illustration several novel waveforms having unusual and useful properties are described.

Most radar design engineers today are acquainted with at least the rudiments of statistical methods and probability concepts. They have studied aspects of detection theory and mastered the operational methods of signal and system analysis. They speak knowingly of "matched filters" and "uncertainty principles." But often this knowledge is fragmentary--quite useful for radar analysis, but fundamentally inadequate for radar synthesis. What is often missing is a sense of perspective, an appreciation of the relative importance of and the interconnections between isolated bits of knowledge--in a phrase, a radar detection philosophy. The rather ambitious purof this paper is to attempt to state such a philosophy--or better a part of such a philosophy since we shall ignore many aspects of the radar detection problem. Specifically we shall not even mention many practical questions such as implementation, effect of system instabilities, approximations, and distortions, countermeasures, etc. Moreover, we shall for various reasons to be discussed consider only "search" type radar applications.

Our intention, thus, is to discuss a theory. Now, the ultimate purpose of any theory in applied science is always to achieve some type of synthesis, i.e., to make it possible to reason back from effects to causes, or from desired performance to system parameters. To be successful, then, our theory must meet three conditions:

1. The model on which the theory is based must at least approximately represent the actual physical situation;
2. The theory must yield a fundamental, complete, and consistent set of parameters and concepts in terms of which both the desired performance and the radar system can be uniquely specified;
3. The theory must include all upper bounds, limiting relationships, or realizability conditions, which prevent the simultaneous achievement of an arbitrary set of parameters.

These three points constitute a rough outline of this paper. More specifically we shall first postulate a model of both the target situation and the radar. We shall then consider two very restricted special cases corresponding to a single target with discrete parameter distributions. Despite the restrictions, a discussion of these cases will lead to quite general statements about those parameters and limiting conditions which relate to the general question of reliability of detection. Next we shall pass to the case of continuous parameter distributions and consider the questions of accuracy and ambiguity and their relation to what might be called the Radar Uncertainty Principle. This section will be illustrated by a number of examples, some of which are rather novel. Finally we shall consider briefly and qualitatively the case of multiple targets and the question of resolution.

---

* The research in this document was supported jointly by the Army, Navy, and Air Force under contract with the Massachusetts Institute of Technology.

1.0 The Radar Model

We shall begin with the postulation of a model. For detection purposes we are principally interested in the received waveform and the way in which it is related to the parameters of the targets and the transmitted waveform. We can avoid a number of trivial steps in the argument if we choose our postulates so as to define directly the received waveform. Hence, we shall assume that:

1. The volume examined by the radar contains a number of point scatterers whose individual properties can be completely described by
   a. the amplitude of the return from each,
   b. the range of the scatterer or delay in the echo,
   c. the velocity of the scatterer or Doppler shift in the echo;
2. The effective duration of the echo from each scatterer is limited, known and independent of the properties of the scatterer;
3. The amplitude of the echo and the velocity of the scatterer are constants, at least during the effective echo duration.

A fourth assumption is required to specify the actual shape of the received waveform, but since this is primarily a matter of nomenclature and requires some development we shall postpone it for the moment. It is easy to raise questions about the necessity, rationality, and implications of the assumptions listed above. Nevertheless, we believe that they are the simplest set of constraints which preserve, at least in some rudimentary form, the major aspects of the radar problem. Moreover, they are the most common assumptions, implied if not expressed, in most discussions of the radar problem, and they have in general the pragmatic justification of leading to mathematical and philosophical problems which are, at least in principle, amenable to solution. Of course, any set of assumptions of this type limits the applicability of the theory to follow. In this respect the second assumption is perhaps the most damaging since it effectively limits the theory to "search" as opposed to "track" applications. Actually there is a rather important point of philosophy involved here about which we shall have more to say later on. Moreover, an additional implication of this second postulate is that we must effectively assume either that the antenna is step-scanned or that the angular coordinates of the target are known a priori. In general we plan to ignore the problem of measuring the angular coordinates of the target. There is nothing particularly fundamental about this and the theory can be modified to include angular measurements with, of course, an increase in complexity. The first and third assumptions are, perhaps, less controversial, and essentially amount to requiring that the observation interval or duration of the echo must be sufficiently short so that acceleration and rapid scintillation effects can be ignored.

Returning now to the fourth assumption, we desire to introduce some symbolic notation for the received waveform and its relation to the transmitted waveform and the target parameters. The first three assumptions permit us to represent the echo signal received at any moment in terms of a canonic signal[1]

$$s_0(t) = Re\left[S_0(t)\, e^{j\omega_0 t}\right] \qquad (1\text{-}1)$$

where

$$S_0(t) = |S_0(t)|\, e^{j\varphi(t)}$$

ω_0 = carrier frequency

so that

$$s_0(t) = |S_0(t)| \cos(\omega_0 t + \varphi(t)) \qquad (1\text{-}1a)$$

The additional assumption will be made that $|S_0(t)|$ and $\varphi(t)$ vary slowly compared with $\omega_0 t$ so that the usual narrow-band assumptions will be justified. For example, it is conveneint to normalize the energy level of $s_0(t)$ by specifying that [2]

$$\int_{-\infty}^{\infty} s_0^2(t)\,dt \approx \frac{1}{2}\int_{-\infty}^{\infty} |S_0(t)|^2\,dt = 1 \qquad (1\text{-}2)$$

Physically, $s_0(t)$ will be interpreted as the return from a "unit" fixed target at zero range. Consequently, except for the effect of antenna scanning or equivalent means of limiting the echo duration, $|S_0(t)|$ and $\varphi(t)$ can be considered as the envelope and phase variation of the transmitted waveform, and hence along with ω_0 are assumed known to the receiver. Thus, using the narrow band assumption, the echo from a target having a range delay τ and a Doppler frequency shift ω rad/sec, $\omega \ll \omega_0$, will be represented as

$$s_\omega(t-\tau) = Re\left[A S_\omega(t-\tau)\, e^{j\omega_0 t}\right] \qquad (1\text{-}3)$$

where

$$S_\omega(t) = |S_0(t)|\, e^{j[\omega t + \varphi(t)]}$$

so that

$$s_\omega(t-\tau) = \qquad (1\text{-}3a)$$

$$= A|S_0(t-\tau)| \cos\left[(\omega_0+\omega)(t-\tau) + \varphi(t-\tau)\right]$$

---

1. Complex notation is employed for simplicity at a later stage and should not be considered in any sense fundamental or mysterious.

2. Although the limits of integration in (1-2), as in other similar integrals to follow, are given as $-\infty$ and ∞, it should be remembered that, in accordance with the second basic assumption, $s_0(t)$ is assumed to be zero outside some relatively short interval of time.

and

$$\int_{-\infty}^{\infty} s_\omega^*(t-\tau)\,dt = A^2 \qquad (1\text{-}3b)$$

= echo signal energy.

Finally our fourth assumption is that the total received signal can be represented as

$$r(t) = n(t) + \sum_i s_{\omega_i}(t-\tau_i)$$

$$= n(t) + \sum_i R_e\left[A_i\, s'_{\omega_i}(t-\tau_i)\, e^{j\omega_0(t-\tau)}\right] \qquad (1\text{-}4)$$

where $n(t)$ represents essentially white Gaussian noise of known power density, N_0 watts/cps (double-sided spectrum)[1]

2.0 The A Posteriori Probability [1]

Given $r(t)$ our problem roughly is to determine the number and parameters of the targets which are present. Of course we cannot expect to do this with absolute certainty because of the noise, if for no other reason. The best that we can hope to achieve is to be able to attach a probability to the truth of any proposition made about the target situation. Specifically we shall assert that a specification for each possible set of values of A, τ, and ω of the probability that there is a target present with those parameters constitutes a complete statement of our knowledge about the target situation obtained from $r(t)$. Considered as a function of A, τ, and ω, this is called the a posteriori (i.e., after reception of $r(t)$) probability distribution which we shall denote by $P(A,\tau,\omega)$.

However, as soon as we try to compute $P(A,\tau,\omega)$ for some particular $r(t)$, a difficult appears. This is, of course, the fact that $P(A,\tau,\omega)$ for each particular A, τ, and ω depends on our expectation prior to receiving $r(t)$ that a target with that particular set of parameters would be present, i.e., $P(A,\tau,\omega)$ depends on the a priori probability distribution $P_0(A,\tau,\omega)$. Other things being equal, the more likely a target is before receiving $r(t)$, the more likely it is afterwards. Specifically it can be shown that if $n(t)$ is white Gaussian noise and if we know, for example, that there is at most one target present, then the a posteriori probability that a target is present and has the particular parameters A, τ, and ω is

$$P(A,\tau,\omega) = \qquad (2\text{-}1)$$

$$= k P_0(A,\tau,\omega)\, e^{\frac{2}{N_0}\int_{-\infty}^{\infty} r(t)\, s_\omega(t-\tau)\,dt}\, e^{-\frac{A^2}{N_0}}$$

where k = normalization constant, independent of A, τ, and ω, so chosen that the sum or integral of $P(A,\tau,\omega)$ over all possible values (including $A=0$) of the parameters is 1.

The difficulty, of course, is that in many cases the radar designer--and indeed the radar user--have little or no knowledge about $P_0(A,\tau,\omega)$. To quote an example of Woodward's, what after all is the "a priori probability of observing an aircraft on a given radar set at a range of ten miles at nine o'clock tomorrow morning?" It is not even clear that this question has any meaning in the sense of mathematical as opposed to subjective probability, since an ensemble of like situations is hard to imagine.

It has been argued with some justification, that, from the the point of view of radar design at least, the dependence of $P(A,\tau,\omega)$ on $P_0(A,\tau,\omega)$ is not very important. The argument essentially depends upon two observations:

1. The only way in which the particular signal received adds to our knowledge of any attribute of the situation is through the integral

$$\Lambda = \int_{-\infty}^{\infty} r(t)\, s_\omega(t-\tau)\,dt \qquad (2\text{-}2)$$

Thus any receiver which computes the integral (2-2) for all possible values of τ and ω is preserving all of that information in $r(t)$ relevant to any decisions about the presence or parameters of an echo signal. Furthermore, the output of this receiver is just sufficient in that any further operations on $r(t)$ will either destroy useful information or imply assumptions concerning the form of $P_0(A,\tau,\omega)$. It is in this sense at least that a receiver performing the cross-correlation type operation (2-2) may be said to be optimum--and the structure of this receiver does not depend on $P_0(A,\tau,\omega)$.

2. The way in which $P_0(A,\tau,\omega)$ enters the computation of $P(A,\tau,\omega)$ is purely as a scale factor or weighting. Thus its influence on the equipment design is essentially that of a gain control adjustment which need not greatly worry the radar designer since it can be left to the user to set this control in accordance with the situation and his own particular prejudices.

At best it seems to us this argument amounts to ducking the issue. There are two reasons why we cannot get rid of the a priori probability problem so easily:

---

1. Contrary to the more usual practice we shall employ throughout a double-sided spectrum, i.e., including both positive and negative frequencies. Hence, for example, the noise power output of a filter with frequency response $H(\omega)$ will be

$$\frac{N_0}{2\pi}\int_{-\infty}^{\infty} |H(\omega)|^2\,d\omega.$$

For a receiver with noise figure $= F$, N_0 is given by $N_0 = \frac{1}{2}kT$, k = Boltzmann's constant, T = absolute temperature.

[1]. Superscripts in brackets refer to the bibliography.

1. Sooner or later in every system design it is necessary to make decisions, e.g., to procede from noncommital probabilities to firm statements right or wrong, that targets are located there and there and there. Such decisions are the rightful concern of the radar designer, if for no other reason than that he may very well be called upon to make them. But as soon as decisions are required the a priori difficulty comes back to haunt us--mixed up now with another disturbing subject, the question of risks or value judgments. It is idle to pretend that decisions based on maximum likelihood, Neyman-Pearson, or minimax criteria avoid the difficult since the selection of such a criterion really amounts in effect to an an assumption about the form of $P_o(R,\tau,\omega)$.

2. The values of $P(R,\tau,\omega)$ for all possible sets of R,τ and ω represent quite a lot of data--more, indeed, than the customer may wish or can assimilate. Some of these parameters, e.g., R, may not contribute much information and the customer may very well suggest that the designer get rid of them by integrating $P(R,\tau,\omega)$ over these parameters. In this event the necessity for knowing or assuming $P_o(R,\tau,\omega)$ is completely unavoidable.

Throughout the remainder of this paper we shall assume that $P_o(R,\tau,\omega)$ is known or has been more or less arbitrarily selected. This really constitutes, of course, a fifth postulate and should perhaps be listed in the preceding section. In the last analysis the justification for this assumption lies in the fact that we are dealing with a theory. In the modern axiomatic sense of the word a theory cannot be tested on the basis of its truth, but only as to its utility. The usefulness of the present theory, including the assumption of known a priori probabilities, has, we believe, been demonstrated many times.

In certain cases the required assumptions concerning the a priori distributions are relatively innocuous. For example, suppose again that there is known to be only one target present and that A^2/N_o is known to be reasonably large compared with 1. Then Λ as defined in (2-2) will, as a result of the narrow-band assumption, be an almost periodic function of τ in the vicinity of the true values of τ and ω, and, indeed, will be almost a sine wave of frequency ω_o and amplitude A^2/N_o. The corresponding period, expressed in terms of range, is one-half the wave length at the frequency $\omega_o/2\pi$, i.e., in most cases a few feet or less. Appealing to a sort of general principle of continuity, it is certainly reasonable to assume that $P_o(R,\tau,\omega)$ is essentially constant over variations of τ corresponding to as small a range difference as a few feet. Thus $P(R,\tau,\omega)$ considered as a function of τ can be expected to alternate rapidly between a large and a small value. Physically the implication is that the range of the target can perhaps be determined quite "accurately" (small fraction of a wave length), but that this determination is highly "ambiguous" (multiples of half-wavelengths). As a result of this ambiguity the fine structure of $P(R,\tau,\omega)$ is essentially meaningless in most cases[1] and the logical operation to perform is to replace $P(R,\tau,\omega)$ by local sums of $P(R,\tau,\omega)$ over half-wavelength intervals, assuming ranges within this interval equally likely. Taking advantage of the narrow-band assumption, this operation is readily carried out by treating the $\omega\tau$ term in $s_\omega(t-\tau)$ as an independent random phase angle, θ, assumed to be uniformly distributed $0\le\theta<2\pi$. Averaging over θ we obtain

$$\bar{P}(R,\tau,\omega) = k P_o(R,\tau,\omega)\, e^{-\frac{A^2}{2N_o}} I_o\left(\frac{A\Lambda_1}{N_o}\right) \qquad (2\text{-}3)$$

where

$$\Lambda_1 = \left| \int_{-\infty}^{\infty} r(t)\, s_\omega^*(t-\tau)\, e^{j\omega t}\, dt \right| \qquad (2\text{-}4)$$

and I_o is the Bessel function of imaginary argument. $\bar{P}(R,\tau,\omega)$ can be thought of as essentially the "envelope" of $P(R,\tau,\omega)$ given by (2-1). More precisely $I_o(x)$ is a monotonic function of x so that Λ_1 plays the same role of a sufficient statistic with respect to $\bar{P}(R,\tau,\omega)$ as Λ plays with respect to $P(R,\tau,\omega)$. Now it is easy to show that Λ_1, considered as a function of τ may be interpreted physically[2] (in the case at least when $\varphi(t)=0$) as precisely the envelope of Λ, which is certainly an intuitively satisfying result considering the assumptions.

---

1. But not in all cases, e.g., not if the radar being considered is one station of an interferometer system.

2. A touch of reality can be given to this "physical" interpretation by considering one of the ways by which Λ and Λ_1, can be computed in practice. Let the interval in which $s_\omega(t)$ is non-zero be $0\le t\le T$. Then Λ can be written

$$\Lambda = \int_{-\infty}^{T+\tau} r(t)\, s_\omega(t-\tau)\, dt$$

which can be interpreted as the output at time $T+\tau$ of a linear filter whose input is $r(t)$ and whose impulse response, $h(t)$, is given by

$$h(t) = s_\omega(T-t)$$

This is the matched filter for this waveform. Clearly the output of this filter as a function of time is precisely equal to Λ for various values of τ. If $s_\omega(t)$ is a narrow-band waveform, then $\Lambda(\tau)$ will also have the appearance of a narrow-band waveform. $\Lambda_1(\tau)$ is precisely the ordinary physical envelope of this waveform, i.e., can be obtained by following the matched filter with an envelope detector.

In the literature $\bar{P}(A,\tau,\omega)$ has in general been called the a posteriori probability in the "random-phase case." This is an unfortunate name since it has led to considerable confusion with the already rather confused question of "coherent" vs. "incoherent" radar. We have gone through the argument leading to $\bar{P}(A,\tau,\omega)$ in some detail in the hope of pointing out that there is really no connection between these two ideas. The receiver computing Λ_1 is every bit as coherent as that computing Λ in the sense that complete knowledge of the internal phase structure of the expected received signal is assumed in each case--it is only the initial phase or detailed local range which is assumed random and equivocal in Λ_1. And, of course, in neither case is there any necessity that the transmitted signal have some regular predictable phase structure ("coherence pulse-to-pulse")--it is merely necessary in each case that the receiver be aware a posteriori of what was indeed transmitted, and this is a condition which, at least in principle, can always be satisfied in radar.

3.0 Simple Detection Situations [2]

In the preceding sections we have discussed the radar model which we have selected and the role which the a posteriori probability plays as a complete measure of our after-the-fact knowledge. But, although the a posteriori probability, supplemented perhaps with some decision method, represents a more or less complete solution to the analysis problem, we must go further before we can do radar synthesis. In particular we must consider the quality of our a posterior knowledge and the way it depends on the various system parameters. There are many sorts of quality judgments which might be applied. We propose to consider just four:

1. The reliability of the detection or determination that a target echo is "there,"
2. The accuracy with which the parameters of the target echo can be measured,
3. The possibility of ambiguities in the determination of target echo parameters,
4. The extent to which two target echoes present simultaneously or overlapping can be resolved and measured separately.

Of these four, the first--reliability of detection--clearly underlies or precedes the other three. In order to acquire some feeling for the detection problem we shall first consider several situations in which the a priori knowledge is, by assumption, such that the other three quality judgments do not apply.

3.1 The Canonic Detection Problem

As we have mentioned before, the question of detection or decision brings up the problem of value judgments, i.e., the relative "costs" to be assigned to the various ways and degrees of being wrong. Fortunately there are several simple situations from which it is possible to draw remarkably general conclusions of great power and utility without having to get deeply involved in such a slippery subject as value judgments and decision criteria. The simplest of these is philosophically almost trivial and might be called the canonic detection problem. We assume that it is known a priori that only one of two possible target situations can ever occur--either there is no target present at all so that the received signal consists of noise alone, or else, one particular known target is present so that the received signal consists of a known echo signal (i.e., known waveform, $A, \tau,$ and ω plus noise.[1] In this case there are only two a posteriori probabilities[2] of interest-- $\bar{P}(A,\tau,\omega)$ and $\bar{P}(0,0,\omega) = 1-\bar{P}(A,\tau,\omega)$ Since our complete a posteriori knowledge of the situation is thus specified by a single number, $\bar{P}(A,\tau,\omega)$, it is clear that the only rational decision process is a comparison of $\bar{P}(A,\tau,\omega)$ with a threshold--announcing desired signal present if $\bar{P}(A,\tau,\omega)$ exceeds this threshold, and otherwise absent. Moreoever, since $I_0(x)$ is monotonic in x, a completely equivalent process, and one which is perhaps more easily interpreted, is merely a comparison of Λ_1 with a different threshold, γ. The choice of γ, of course, depends on the specified values of $A, \tau, \omega, P_0(A,\tau,\omega)$, and on the appropriate value judgments selected to rate the performance of the decision process. But this is the only way in which the questions of either the value judgments or the a priori probability enter the problem. Thus the significance of the a posteriori approach, in this case at least, is that we can state quite unequivocally that the form of the optimum decision process, i.e., compare Λ_1 with a threshold, will not depend on the particular value judgment chosen, which is really a quite remarkable and important conclusion. Indeed if at least some relative degree of invariance to such an emotional quantity as value judgments were not obtained in as simple a decision problem as this, we would have very serious doubts about the likelihood of any really general and useful conclusions coming out of the present approach.

The remaining result of interest in the canonic detection problem is a determination of those attributes of the received echo waveform which influence the decision performance. First, it is necessary to point out that the performance of the detector in this simple problem can be completely characterized by two conditional probabilities:

P_d = probability of announcing echo signal present if there actually is such a signal present = Probability of Detection,
P_f = probability of announcing echo signal present if there actually is not a signal present = Probability of False Alarm.

1. Clearly such a situation is almost too trivial to ever be representative of an actual radar problem. Nevertheless, certain communications problems, e.g., synchronous PCM, are represented rather accurately by this model.

2. Throughout this paper we shall assume that we are dealing with the "random-phase case" so that $\bar{P}(A,\tau,\omega)$ rather than $P(A,\tau,\omega)$ is the appropriate probability distribution.

It is easy to show by an analysis of the statistical properties of Λ, that P_d and P_f are functions of just two parameters:

$$R = \frac{A^2}{N_o} = \frac{\text{echo signal energy}}{\text{effective input noise power/cps.}}$$

$$\frac{\gamma^2}{N_o} = \frac{(\text{threshold voltage})^2}{\text{effective input noise power/cps.}}$$

The parameter γ^2/N_o can be eliminated and P_d plotted as a function of P_f with R as a parameter. The resulting family of curves (Fig. 1) has been called the receiver detection characteristic.[2] In accordance with the argument of the preceding paragraph the interpretation to be put on these curves is the following. For a given R any pair of values of P_d and P_f lying on the corresponding curve can be obtained by comparing Λ with an appropriate threshold. The best operating point is, of course, a function of the selected value judgment and, in general, $P_o(A, \tau, \omega)$. But in any case the performance so obtained is optimum in the sense that no pair of values of P_d and P_f above or to the left of this curve can be obtained by any means with the given R. Any actual receiver which fails to compute Λ, or its equivalent will yield operating points lying below this curve[1].

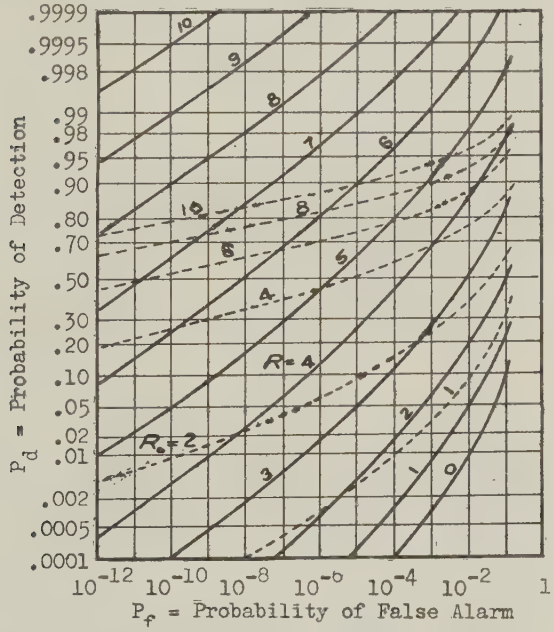


Fig. 1: Receiver Detection Characteristics

We have found the curves of Fig. 1 very useful for computing the performance of radars and other systems (as compared, for example, with the rather nebulous empirical rule that "reliable" detection requires some signal-to-noise ratio to be greater than 1).[1] But for our present purposes the most important conclusion of the last paragraph is that the detector performance, in the canonic detection situation at least and in so far as it depends on the actual signal received, depends only on the ratio of the desired echo signal energy to the noise power per cycle per second, and not upon any other attribute of the waveform (e.g., bandwidth, waveshape, etc.). That the performance should in principle depend on the ratio, R and not alone on the more common signal-to-noise power ratio is certainly not surprising in view of the Equipartition Law of physics. But when translated into other terms, e.g., the observation that a pulse radar and a CW radar will have the same detection performance on a given target for the same average received power and observation time, despite the large difference in bandwidth, the argument is not always so readily believed.

Before going on to consider more complicated decision problems it is worthwhile to investigate the effect of an unknown amplitude,[2] A, on what is otherwise the canonic decision problem. There are, perhaps, two extreme situations--one in which it is desired to estimate A as well as detect the presence of the signal, and the other in which the actual value of A contributes essentially no information and only a yes-or-no answer about presence is desired. The first situation brings up the question of accuracy; indeed, our discussion here will serve as the prototype for later discussions with respect to τ and ω. The second situation provides another example of the proper way to handle a "stray" or non-information-carrying parameter. In the second case, in particular, it is necessary to make an assumption about the a priori distribution of amplitudes. For purposes of illustration we shall choose the Rayleigh distribution which in many cases is a reasonable approximation to the actual distribution and has the advantage of being easy to manipulate. That is we shall assume that

$$P_o(A, \tau, \omega) = P_o(\tau, \omega) \frac{A}{A_o^2} e^{-\frac{A}{2A_o^2}} dA \qquad (3.1\text{-}1)$$

If only a yes-no answer about the presence of a target is desired, then paralleling the random-phase arguments what we must compute is

$$\bar{P}(\tau, \omega) = \int_0^\infty \bar{P}(A, \tau, \omega)\, dA \qquad (3.1\text{-}2)$$

1. It is important to point out, however, in view of the many arguments which have arisen over the past few years, that a receiver which can be considered as only in the grossest sense computing Λ, may often (indeed one is tempted to say almost always) have an operating point only slightly below (equivalent to a few db change in R) the optimum. In other words the optimum represented by comparing Λ, with a threshold is, in an operational sense, very broad. This is another example of the relative invariance which we consider such a satisfying feature of the theory

1. Indeed we consider the most meaningful and fundamental form of the radar equation to be that which relates R, rather than received signal power, to transmitted power, antenna gain, range, etc.
2. It should be remembered that we are still assuming that, whatever the value of A, it remains constant for the duration of the signal. In physical terms we are thus modeling the slowly-scintillating target only. Rapidly-scintillating targets present much more difficult problems.

i.e., the sum of the a posteriori probabilities for each value of A. The integration can be easily carried out and it can be shown that a comparison of $p(\tau,\omega)$ with a threshold is completely equivalent to comparing Λ_1 with a threshold as before. The difference from the case of known amplitude comes in the values of P_d and P_f. The receiver detection characteristic is readily computed and the resulting family of curves is shown dotted in Fig. 1. The parameter in this case is the ratio $R_0 = A_0^2/N_0$, where A_0 is the most probable target echo amplitude. The most significant attributes of these curves are the much lower value of P_d resulting for high values of R_0 compared with the non-fluctuating case for the same value of R, and the effective saturation in P_d accompanying an attempt to increase P_d by increasing R_0. These results have an important influence on radar design, but further discussion of these effects is outside the scope of this paper.

Suppose, however, it is necessary to determine the actual amplitude, say A_1, of the target.[1] The question then arises as to the accuracy with which A_1 can be measured. The a posteriori probability of a particular value, A, is, from (2-3)

$$p(A,\tau,\omega) = k R(A,\tau,\omega)\, e^{-\frac{A^2}{2N_0}}\, I_0\left(\frac{A\Lambda_1}{N_0}\right) \qquad (3.1\text{-}3)$$

Now as we have already shown the ratio A_1^2/N_0 must be large ($\gg 1$) if the signal is to be detected reliably. Or alternately we shall show that A_1^2/N_0 must be large if A_1 is to be determined accurately. From either point of view we conclude that the interesting range of A in (3.1-3) is the neighborhood of A_1 and that A^2/N_0 and $A\Lambda_1/N_0$ will both be $\gg 1$. If $R_0(A,\tau,\omega)$ is continuous, we are probably entitled to assume that the variation of $R_0(A,\tau,\omega)$ in the neighborhood of A_1 will be small so that the a priori probability may be effectively included for present purposes in the constant k. $p(A,\tau,\omega)$ presumably has a large peak at $A \approx A_1$; the precise location of the maximum, i.e., the most probable value of A, is determined by

$$\frac{\partial\left(p(A,\tau,\omega)\right)}{\partial A} = 0 \qquad (3.1\text{-}4)$$

or

$$A I_0\left(\frac{A\Lambda_1}{N_0}\right) - \Lambda_1 I_1\left(\frac{A\Lambda_1}{N_0}\right) = 0 \qquad (3.1\text{-}5)$$

Using the fact that $\frac{A\Lambda_1}{N_0} \gg 1$ and preserving only the first terms of the asymptotic expansions

$$I_0(x) \sim \frac{e^x}{\sqrt{2\pi x}}\left(1 + \frac{1}{8x} + \cdots\right) \qquad (3.1\text{-}6)$$

$$I_1(x) \sim \frac{e^x}{\sqrt{2\pi x}}\left(1 - \frac{3}{8x} + \cdots\right), \qquad (3.1\text{-}7)$$

the solution of (3.1-5) is simply

$$A = \Lambda_1 \qquad (3.1\text{-}8)$$

with an error the order of $\frac{N_0}{\Lambda_1} \frac{1}{2}$. Equation (3.1-8) is certainly reasonable. In particular we note that as $N_0 \to 0$, $\Lambda_1 \to A_1$.

We now wish to focus attention on a series of cases in which the actual amplitude is A_1, and to consider the distribution of values of A which would result from (3.1-8). A study of the form of Λ_1 (see (2-4) and footnote 1, page 2) shows that Λ_1 has the same distribution as the envelope of a sinusoid of amplitude A_1 plus a narrow-band Gaussian noise with variance N_0. This problem has been considered by Rice,[2] who has shown that the distribution for A is approximately normal (for $A_1/N_0 \gg 1$) with mean value equal to A_1, as we should expect, and variance equal to N_0. Thus the normalized effective standard deviation is approximately

$$\frac{\sigma_A}{A} = \frac{1}{\sqrt{R}} \qquad (3.1\text{-}9)$$

which is the result we sought. Typically R is the order of 100 or more so that a relative accuracy of better than 10 per cent in the determination of A_1 is reasonable.

Completely aside from the various approximations employed, the procedure of the previous paragraph can be questioned on philosophical grounds. In outline what we did was the following:

1. Determined that value of A, say A', for which $p(A,\tau,\omega)$ had its maximum in the vicinity of $A = A_1$;
2. Found the probability distribution of A' over all the received signals for which $A = A_1$.

The difficulty with this approach is that we have specified the form of the estimation operation in advance, i.e., choose that value of A which maximizes $p(A,\tau,\omega)$. This is certainly a reasonable thing to do, but there are other possible reasonable operations, e.g., choose that value which will minimize $(A - A_1)^2$ on the average. We cannot be sure a priori that the values of A selected on the basis of different criteria might not be different and have different error distributions.

Fortunately there is another approach which gets around this difficulty and, in addition, is both easier to carry through and more illuminating. The difference, essentially, is that instead of computing the distribution of some arbitrarily selected quantity such as $A' = \max(p(A,\tau,\omega))$ over all received signals with fixed A_1, we shall compute the distribution of values of A which might have led to a particular $r(t)$. But this is precisely the a posteriori probability $p(A,\tau,\omega)$ for

1. Situations in which the amplitude of the return is a useful piece of information appear to be rare in radar systems. The commonest example, perhaps, is in monopulse systems, and here it is not so much the amplitude of the echo signal for a single radar as it is effectively the ratio of the amplitudes for two radars which matters. It seems possible that future radar systems may perhaps find amplitude information useful as an aid to identification.

this particular $r(t)$. Now, of course, $P(A,\tau,\omega)$ can tell us something general about the accuracy of estimating A_1 only if $P(A,\tau,\omega)$, in the vicinity of $A=A_1$ at least, has essentially the same shape for nearly all received signals, $r(t)$, having the same A_1. But this basically implies that we can measure A_1 with high accuracy which is exactly the condition we wish to achieve and are most interested in. Hence, the argument is self-consistent--if $P(A,\tau,\omega)$ has a large spike in the neighborhood of some particular A, i.e., if it is highly probable that a signal with A_1 in this neighborhood is present, then the shape of $P(A,\tau,\omega)$ in this neighborhood describes the accuracy with which we can determine A_1 since $P(A,\tau,\omega)$ is exactly the probability that some A other than A_1 could have been present in $r(t)$. In such a case $P(A,\tau,\omega)$ will be determined almost entirely by the echo signal and will be almost independent of the particular noise present so that we may make general statements. If $P(A,\tau,\omega)$ does not have a large isolated spike then, although $P(A,\tau,\omega)$ still measures the distribution of possible values of A_1 that might have produced the particular $r(t)$ received, the accuracy is presumably low and no general statements can be made since the noise has had a major effect on $P(A,\tau,\omega)$. This same argument applies to any other parameter as well as to A and is the one we shall employ for τ and ω.

For the particular problem being considered $P(A,\tau,\omega)$ is given by (3.1-3). As before we shall assume that $P_0(A,\tau,\omega)$ is essentially constant over the interesting neighborhood so that

$$P(A,\tau,\omega) \sim e^{-\frac{A^2}{2N_0}} I_0\left(\frac{A\Lambda_1}{N_0}\right) \qquad (3.1\text{-}10)$$

Now the function Λ_1 can be broken up into two terms by writing $r(t) = n(t) + s_\omega(t-\tau)$;

$$\Lambda_1 = \left| \int_{-\infty}^{\infty} s_\omega(t-\tau) S^*_\omega(t-\tau) e^{j\omega t} dt + \right.$$

$$\left. + \int_{-\infty}^{\infty} n(t) S^*_\omega(t-\tau) e^{j\omega t} dt \right| \qquad (3.1\text{-}11)$$

The first integral is equal to A_1 = amplitude of echo signal present. The supposition that $P(A,\tau,\omega)$ has a large spike in the neighborhood being considered implies that the second term is with high probability small and that A^2/N_0 is large so that

$$P(A,\tau,\omega) \sim e^{-\frac{A^2}{2N_0}} I_0\left(\frac{AA_1}{N_0}\right) \qquad (3.1\text{-}12)$$

$$\sim e^{-\frac{(A-A_1)^2}{2N_0}} \qquad (3.1\text{-}13)$$

where we have replaced $I_0\left(\frac{AA_1}{N_0}\right)$ by the first term of its asymptiotic expansion (3.1-6) and completed the square. Thus $P(A,\tau,\omega)$ is approximately normally distributed near A_1 with mean = max $[P(A,\tau,\omega)] = A_1$ and standard deviation $\sigma_A = \sqrt{V_A} = A/\sqrt{R}$ as before. In addition to being simpler than the first approach, this method has, in Woodward's words "the remarkable feature----that [accuracy] can be studied in the absence of noise" since the effect of our argument was to remove the noise term from Λ_1 in so far as computing $P(A,\tau,\omega)$ was concerned.

In the preceding paragraphs we have considered nearly all the questions which can arise in the canonic detection problem, with the exception of those which more properly belong to the study of values and which determine the actual value of the threshold to be employed. To summarize we have observed that:

1. For the canonic detection problem the optimum form of decision process is a comparison of the envelope of the output of a matched filter or cross-correlator with a threshold. The form of this decision process is independent of either the type of value judgment selected or the a priori probabilities;
2. The reliability of detection, in so far as it depends on the characteristics of the radar and the target, depends only on the ratio of received signal energy to noise power per cycle per second, i.e., on R;
3. The accuracy with which the parameter A can be determined is measured by the standard deviation of $P(A,\tau,\omega)$. For $R>>1$ (and we observed that R has to be much greater than one if the detection performance is to be reliable) the shape of $P(A,\tau,\omega)$ in the neighborhood of the correct value of A is nearly the same for all $r(t)$ and is essentially equal to $P(A,\tau,\omega)$ with $s_\omega(t-\tau)$ substituted for $r(t)$. Specifically the standard deviation of A is given by

$$\sigma_A = \frac{A}{\sqrt{R}}$$

3.2 The Case of M Orthogonal Signals [2]

We now wish to consider another simple detection problem which is somewhat more directly related to the radar problem than that discussed in the preceding section. As before we shall assume that at most one target echo signal is present at any one time or during any one observation interval. We shall further assume that the amplitude, A, is known a priori and that the initial phase angle is random and informationless. However, unlike the preceding case we shall assume that the parameters τ and ω of the echo signal are not known a priori but instead any one of M signals of the form $As_{\omega_i}(t-\tau_i)$ may be present with equal probability. We shall further assume that these M signals are mutually orthogonal, i.e., that

$$\int_{-\infty}^{\infty} S_{\omega_\mu}(t-\tau_\mu) S^*_{\omega_j}(t-\tau_j)\, dt = \begin{cases} 2; & \mu = j \\ 0; & \mu \neq j \end{cases} \qquad (3.2\text{-}1)$$

where the star indicates comples conjugate. Although this set of assumptions is clearly a closer approximation to actual radar problems than the canonic detection problem in that τ and ω are treated as unknowns, the assumed discrete nature of the a priori distributions in τ and ω, and

the requirements that (3.2-1) be satisfied, are obviously unrealistic. The next section will be largely devoted to removing these restrictions, but their value for the moment is that again we will be able to say something of general value about the reliability of detection without getting involved in the accuracy, ambiguity, and resolution questions. Indeed what we hope eventually to be able to conclude--and this is one of the principle points of our detection philosophy--is that the radar detection problem breaks down into two essentially independent problems. The first of these is to adjust the radar parameters, particularly those which appear in the radar equation, in order to obtain sufficient received energy under the desired conditions to produce reliable detection of the fact that a signal is present. The important point is that the required signal energy, or better the required value of R, can be stated almost independently of the received waveform. The second problem, then is to adjust the received waveform, by means, of course, of choosing the transmitted waveform, in order to obtain the desired performance in terms of accuracy, ambiguity, and resolution.

For the moment, then, we are interested in the reliability of detection if there are M orthogonal signals which might possibly be present one at a time, instead of just one known signal. From the beginning it is apparent that the present problem is philosophically considerably more involved than the simple canonic problem in that the a posteriori probability distribution (which is still given by (2-3) now consists of a set of M numbers[1] instead of just two. In particular it is no longer possible to circumvent more or less completely the question of value judgments--there are many meaningful and different ways in which the question "Is there a signal present?" can be asked. For our present purposes, however, as outlined in the preceding paragraph, it will perhaps suffice to demonstrate the essential invariance of the system behavior by analysing two examples.

The first of these is perhaps the most obvious formulation of the pure detection problem in the present situation. If we reduce the detection problem to a simple binary decision then the performance can be completely described as before by the various probabilities of being right and wrong, such as P_d and P_f. In particular for our first example we seek the optimum receiver detection characteristic relating

P_d = probability of announcing that a (any) signal is present when indeed there is a (some) signal present,

and P_f = probability of announcing that a (any) signal is present when actually only noise is present.

From the discussion in preceding sections it should be obvious that the corresponding optimum decision operation is to compare

$$\sum_{i=1}^{M} p(A, r_i, \omega_i)$$

with a threshold. Using the assumption of orthogonality (which assures statistical independence among the terms of the sum) and assuming that M is large enough so that the central limit theorem of statistics may be employed, it is a more or less straight-forward problem to show that the receiver detection characteristic has essentially the form of the solid lines of Fig. 1 where the parameter, R, is to be interpreted, not as equal to A^2/N_o but rather

$$R \approx A^2/N_o - \ln M \qquad (3.2\text{-}2)$$

Actually the approximations are such that this expression for R is slightly pessimistic.

The most important conclusion to be derived from this result is that the value of A^2/N_o required for a given performance is only logarithmically dependent on M. Indeed we would be justified in stating that to a first approximation the required value of A^2/N_o is independent of M. The essential truth of this statement is perhaps best illustrated by an example. For "reliable" detection (e.g., $P_d = 0.99$, $P_f = 10^{-5}$) the R of Fig. 1 typically must be the order of 50. If $M = 20{,}000$ then A^2/N_o must equal 60 which is an increase of just 0.8 db over the value of A^2/N_o (=50) which would be required for the same P_d and P_f if $M = 1$, i.e., if τ and ω were fixed and known a priori.

As a second example we consider a decision process of a somewhat different nature. We postulate the existence of a receiver with M output channels, one for each possible echo signal. Each channel has two output states corresponding to "signal present" and "signal absent" and we suppose that the channels are so interlocked that only one channel can indicate "signal present" at a time.

We then seek the optimum receiver detection characteristic relating

P_d = probability that a particular signal will be announced if indeed that particular signal is present,[1]

and $$P_f = \sum_{i=1}^{M} P_{f_i}$$

where P_{f_i} = probability that the i^{th} signal will be announced if indeed the i^{th} signal is not present.

Although this criterion is obviously quite a bit stiffer than that discussed in the first example, we cannot argue that the resulting detector will be better or worse than the first detector unless the use to which the data is to be put and the corresponding appropriate value judgments are considered. Our purpose, however, is not to compare these criteria but rather to show that even in this second case a large increase in M requires only a small increase in the value of A^2/N_o to keep the reliability of detection constant. It is easy to argue that the optimum detection process

1. The probability of each of the M signals together with the probability, $p(0,0,0)$ of noise alone.

1. We assume that all signals are treated alike so that P_d is the same for each channel.

in this second case is to compare Λ_i separately with a threshold for each signal, announcing that signal present, if any, for which the corresponding Λ_i is most in excess of the threshold.[1] As a result of the assumed orthogonality among the signals the solid curves of Fig. 1 may then be interpreted as plotting the relationship between P_d and P_f. Assuming that all signals are treated alike

$$P_f = M P_{f_i} \qquad (3.2\text{-}3)$$

To illustrate (using the same example before)--suppose $P_d = 0.99$, $P_f = 10^{-5}$, $M = 20{,}000$. Then $P_{f_i} = 5 \times 10^{-10}$ and the required value of A^2/N_o is $(8.8)^2 = 77.4$ which is a 1.9db increase over the value of A^2/N_o required for τ and ω known *a priori* and a 1.1db increase over the value required in the first example--in neither case a very significant amount considering the size of M.

4.0 Detection in the Case of Continuous Parameter Distributions; Accuracy and Ambiguity.

As soon as we pass from the discrete *a priori* distribution assumed for τ and ω in the preceding sections to more realistic continuous distributions, a whole host of problems arise. These problems are not only of a mathematical or computational nature but also, as we saw in a rather elementary way in the example of section 3.2, involve quite complicated questions of value judgments and problem formulation. Roughly the difficult is that it is no longer possible to state performance measures in black and white terms;, there are various ways or degrees of being wrong and the penalties must be weighted accordingly. Nevertheless, if we have any hopes of evolving a useful theory we must face these problems, even if our conclusions are more qualitative than quantitative.

As before we shall assume that at most one target echo is present during any observation interval and that the initial phase angle of the return is random and informationless. We shall further assume, for simplicity and to be definite, that:

1. The amplitude, A, is known and constant, independent of τ and ω.
2. The *a priori* probability density $p_o(\tau,\omega)$ is a constant for all values of τ and ω lying inside the rectangle in the $\tau\omega$ plane bounded by the lines $\tau = 0$; $\tau = \Theta$, $\omega = \pm\frac{\Omega}{2}$. In other words all pairs of values of τ and ω satisfying the conditions $0 \le \tau \le \Theta$ and $|\omega| \le \frac{\Omega}{2}$ will be assumed equally likely.

Other, less restrictive, assumptions can be handled, at least qualitatively, but these will serve to give the principal outlines of what can be done.

Loosely, we shall be concerned with three questions. The first of these is essentially the same question considered previously, i.e., what is the reliability of detection, where by detection we have in mind essentially the same sort of decision as in the first example of section 3.2. And, indeed, our method of handling this problem will be to replace the actual continuous parameter situation by an approximately equivalent discrete orthogonal problem of the type analysed in section 3.2. The other two questions are new. One is the question of the accuracy with which the parameters τ and ω can be measured once it has been ascertained that an echo is indeed present. We have considered the question of accuracy before with respect to the measurement of amplitude in the canonic detection problem. Essentially the same method will be employed here for τ and ω. The third question concerns the possibility of ambiguity, i.e., are there other values of τ and ω significantly removed from the proper values which might conceivably be misconstrued as the right values. Actually there are two ways in which an ambiguity might arise. One possibility is that the noise accompanying a particular echo might look sufficiently like some other echo that the *a posteriori* probability of this latter signal might be large. This type of ambiguity really has more the character of a false alarm--if the detection is "reliable" than such ambiguities should be rare. On the other hand the structure of the signal may be such that two echoes from different targets may look much alike, e.g., the "second-time-around" target in conventional pulse radar. This is the type of ambiguity we wish to study. The principal objective of our study, of course, is radar synthesis. Hence, with respect to accuracy and ambiguity we shall seek both for those attributes of the radar which influence these problems and for appropriate limit theorems or realizability conditions on the types of accuracy and ambiguity performance which can be obtained. We shall find, of course, that unlike the reliability of detection problem, the important parameters influencing accuracy and ambiguity are those which describe the waveform, and we shall discuss an important performance constraint on the waveform which perhaps deserves the title of the Radar Uncertainty Principle.

As a result of the various assumptions which we have made, the *a posteriori* probability density for the present situation wan be written in a somewhat simpler form than (2-3). Specifically,

$$\bar{p}(\tau,\omega) = k I_o\left(\frac{A}{N_o}\Lambda_i\right) \qquad (4\text{-}1)$$

where we have absorbed into k a multitude of terms including the *a priori* probability density. Now it should be obvious that if we are going to judge the reliability of detection in the present case on the same basis as the first example of section 3.2, i.e., in terms of

P_d = probability of announcing that a (any) signal is present when indeed there is a (some) signal present;

1. The similarity of this process to that carried out in the usual range-gated radar--particularly those of the so-called "predetection integration" type--is perhaps worth pointing out.

and P_f = probability of announcing that a (any) signal is present when actually only noise is present;

then by analogy with that example the optimum detection procedure is to compare

$$\int_0^{\Omega} d\tau \int_{-\Omega/2}^{\Omega/2} \bar{P}(\tau,\omega)\, d\omega \qquad (4\text{-}2)$$

with a threshold. We can even perhaps imagine an infinite number of cross-correlators followed by non-linear weighting devices, a summing circuit, and a comparison circuit which would physically carry out this operation. But a real difficulty arises when we try to compute P_d and P_f since the noise outputs of these individual channels will not in general be independent--the corresponding signals will not in general be orthogonal or uncorrelated.

The question of the way in which the various possible echo signals are correlated is a most important one for our study since it not only influences the value of A^2/N_0 required for a given reliability of detection[1], but also has a major effect on the accuracy and ambiguity question. To see this let us consider two signals--$s_{\omega_1}(t-\tau_1)$ and $s_{\omega_2}(t-\tau_2)$. Let us then compute both $\Lambda_1(\tau_1,\omega_1)$ and $\Lambda_1(\tau_2,\omega_2)$ in the case in which $s_{\omega_1}(t-\tau_1)$ is actually present

$$\Lambda_1(\tau_1,\omega_1) = \left| \int_{-\infty}^{\infty} \left(n(t) + s_{\omega_1}(t-\tau_1)\right) S^*_{\omega_1}(t-\tau_1) e^{j\omega_1 t}\, dt \right|$$

$$= \left| A + \int_{-\infty}^{\infty} n(t)\, S^*_{\omega_1}(t-\tau_1) e^{j\omega_1 t}\, dt \right| \qquad (4.3)$$

$$\Lambda_1(\tau_2,\omega_2) = \left| \int_{-\infty}^{\infty} \left(n(t) + s_{\omega_1}(t-\tau_1)\right) S^*_{\omega_2}(t-\tau_2) e^{j\omega_2 t}\, dt \right|$$

$$= \left| \frac{A}{2} \int_{-\infty}^{\infty} S^*_{\omega_1}(t-\tau_1) S_{\omega_2}(t-\tau_2)\, dt \right.$$

$$\left. + \int_{-\infty}^{\infty} n(t)\, S^*_{\omega_2}(t-\tau_2) e^{j\omega_2 t}\, dt \right| \qquad (4.4)$$

If, as we have argued before, the detection is to be reliable, then the first term in (4-3) must be much larger than the second[2] so that

$$\Lambda_1(\tau_1,\omega_1) \approx A \qquad (4\text{-}5)$$

If in addition $s_{\omega_1}(t-\tau_1)$ and $s_{\omega_2}(t-\tau_2)$ are highly correlated, by which we mean that

$$\left| \frac{1}{2} \int_{-\infty}^{\infty} S^*_{\omega_1}(t-\tau_1) S_{\omega_2}(t-\tau_2)\, dt \right| \approx 1 \qquad (4\text{-}6)$$

then it will also be true that

$$\Lambda_1(\tau_2,\omega_2) \approx A. \qquad (4\text{-}7)$$

Thus $$\Lambda_1(\tau_2,\omega_2) \approx \Lambda_1(\tau_1,\omega_1) \qquad (4\text{-}8)$$

and so $$\bar{P}(\tau_1,\omega_1) \approx \bar{P}(\tau_2,\omega_2). \qquad (4\text{-}9)$$

Whether $\bar{P}(\tau_1,\omega_1)$ will actually be greater than or less than $\bar{P}(\tau_2,\omega_2)$ will depend on the particular noise waveform present, but will not depend very much on which signal is present. In other words, when $s_{\omega_1}(t-\tau_1)$ is actually present, we can not really be sure if it is $s_{\omega_1}(t-\tau_1)$ or $s_{\omega_2}(t-\tau_2)$ which is present. Thus if two possible received waveforms are highly correlated in the sense of (4-6) and one of them is present, the determination of which one is present is fundamentally ambiguous and no amount of clever data processing can resolve this ambiguity.

In a similar situation suppose that for some fixed value of $\omega_2 = \omega_1$ and for all values of τ lying in some interval $\Delta\tau$ centered on τ_1 the corresponding signals $s_{\omega_1}(t-\tau)$ are highly correlated in the sense of (4-6). Then if one of these signals is present we shall not be able to determine which one, i.e., we shall not be able to measure τ_1 with an accuracy greater than the order of $\Delta\tau$. This, of course, is in essence the same argument as we employed in section 3.1 with relation to the accuracy of estimating A.

To be sure the accuracy and ambiguity situations depend not only on the degree of correlation of the various signals but also on the ratio A^2/N_0 --we shall have more to say about this in what follows. But the important point is that the limiting possible performance with respect to accuracy and ambiguity depends not on ingenuity in processing the received signals but rather on the shape of the received signals themselves and in particular on the extent to which the various received signals are correlated. It behooves us, therefore, to study the various possible forms which this correlation can take. Such a study will permit us not only to give more-or-less complete answers to the detectability, accuracy, and ambiguity questions in particular cases but will lead to one of the most important theorems constraining radar performance--the Radar Uncertainty Principle.[1]

4.1 Waveform Examples; Radar Uncertainty Principle

We are interested in the behavior of the function

$$\psi(\tau;\omega') = \frac{1}{2} \left| \int_{-\infty}^{\infty} S^*_{\omega_1}(t-\tau) S_{\omega_2}(t-\tau_2)\, dt \right|$$

$$= \frac{1}{2} \left| \int_{-\infty}^{\infty} S_0(t) S^*_0(t+\tau')\, e^{j\omega' t}\, dt \right| \qquad (4.1\text{-}1)$$

1. As we have anticipated and will show this influence is actually rather small.

2. The mean square values of the real and imaginary parts of the second term in both (4-3) and (4-4) are each equal to N_0.

where

$$\tau' = \tau_2 - \tau_1$$

$$\omega' = \omega_2 - \omega_1$$

and the second expression has been obtained from the first by an elementary change of variable. Physically $\psi(\tau',\omega')$ can be loosely interpreted as the output in the absence of noise of a cross-correlator corresponding to a particular signal when a second signal with a delay less by τ' and a frequency shift less by ω' than that particular signal is present. Alternately $\psi(\tau',\omega')$, considered as a function of τ' and with τ' interpreted as time, is, except for a scale factor and ignoring noise, precisely the time waveform corresponding to the envelope of the output of a matched filter for a signal at a particular frequency when a signal at a frequency ω' less is present. Although $\psi(\tau',\omega')$ has a number of interesting and important mathematical properties,[3] it is perhaps more expedient for our present purposes to procede to a consideration of several examples.

4.11 Impulse or CW Radar

We shall assume for our first example that the received waveform is a single pulsed-sinusoid of constant amplitude and duration, T. Thus, recalling the normalization of equation (1-2)

$$s_0(t) = \begin{cases} \sqrt{2/T}\,; & 0 \le t \le T \\ 0\,; & \text{elsewhere} \end{cases} \qquad (4.11\text{-}1)$$

Physically there are two interesting extreme situations to which this example might correspond

a. CW Radar--where T is essentially the "time on target" and typically

$$T >> \Theta$$

but $$\frac{2\pi}{T} << \Omega$$

b. Impulse Radar--where T is now simply the impulse duration and typically

$$T << \Theta$$

but $$\frac{2\pi}{T} >> \Omega$$

$\psi(\tau',\omega')$ is shown in Fig. 2 plotted as amplitude above the τ'-ω' plane. We note that $\psi(0,0)$ is the highest point in the plane--which is reasonable since $\psi(\tau,\omega)$ is proportional to the output of the cross-correlator when the corresponding signal is present. Indeed we could prove the general result that for any waveform $s_0(t)$ of finite duration

$$\psi(0,0) = 1 > \psi(\tau',\omega') \; ; \quad \tau',\omega' \neq 0 \qquad (4.11\text{-}1a)$$

Philosophically this equation can be interpreted to mean that if the noise is vanishingly small, the parameters of an echo signal can be determined with perfect accuracy and without ambiguity--a physically satisfying conclusion.

An alternative way of representing $\psi(\tau',\omega')$ is shown in Fig. 3 and 4 where we have chosen the τ' and ω' scales so as to more clearly illustrate the difference between CW and Impulse Radar respectively. In these two figures, as in most of the remaining plots of $\psi(\tau',\omega')$ in this paper we have chosen to indicate the magnitude of $\psi(\tau',\omega')$ as the density of shading in the two-dimensional τ'-ω' plane. Moreover, for simplicity we have somewhat arbitrarily restricted the degrees of shading to just three--black corresponding to highly correlated regions, i.e., $\psi(\tau',\omega') \approx 1$, cross-hatch corresponding to weakly correlated regions, i.e., $\psi(\tau',\omega') \approx 0$ and unshaded corresponding to uncorrelated regions, i.e., $\psi(\tau',\omega') = 0$.

Let us now consider the problem of determining the receiver detection characteristic for the case, say, of the CW radar of Fig. 3. We first note that it is not necessary to provide a channel in our detector for every possible pair of values of τ and ω as we assumed earlier. For example, since $\Theta \ll T$ the channel corresponding to $\tau = 0$ and any particular ω will have an output under all circumstances which is very highly correlated, i.e., nearly identical, with the output of the channels for the same ω and any $\tau \le \Theta$. Thus we could replace all these channels in the integration (4-2) with just one channel, say $\tau = 0$. Similarly in frequency except channels separated $2\pi/T$ from a given channel are nearly uncorrelated with the given channel. Thus approximately $\Omega T/2\pi$ channels are required in frequency to cover the expected range of targets and the outputs of these channels are nearly orthogonal. Thus finally the detection performance in the case of a CW radar cannot be very different from that of the first example of section 3.2 with $M = \Omega T/2\pi$. Of course we could easily be off by a factor of 2 or more in this value of an equivalent M but since the required value of A^2/N_0 depends on M only logarithmically such an error has negligible significance. Moreover the whole argument is somewhat academic since, as we showed in section 3.2 the increase in A^2/N_0 required even for a large value of M is quite small. Of course a precisely dual argument holds in the case of the impulse radar, the required value of M being the order of Θ/T.

Next we shall examine the accuracy[1],[4] with which the parameters τ and ω can be determined in the case of a CW radar. Of course, since for the allowed variation in τ, all possible signals are highly correlated, essentially no estimate of τ can be given unless A^2/N_0 is very large, as is physically obvious. Measurements of ω are more interesting. Using the same argument as in section 3.1 the procedure is to identify the variance in the measurement of ω with the mean square width of the spike in $\psi(\tau,\omega)$ computed in the absence of noise. As should be readily apparent this is precisely the same as the mean square width of the spike at the origin of $I_0\left(\frac{A^2}{N_0}\psi(0,\omega)\right)$. Making the usual approximations we obtain

$$\sigma_\omega \approx \frac{2\sqrt{3}}{T\sqrt{A^2/N_0}} \qquad (4.11\text{-}2)$$

as the standard deviation of the error in ω assuming $A^2/N_0 \gg 1$. If we take $2\pi/T$ as the width in frequency of the central spike in $\psi(\tau,\omega)$ and if we assume $A^2/N_0 = 50$ then (4.11-2) states that we should be able to determine ω to about 1/10 of this width. This is somewhat better than we are able to achieve in practice because of the effect of systematic errors which have been ignored. Equation (4.11-2) is a special case of a general result which states that

$$\sigma_\omega \approx \frac{1}{\sqrt{\frac{A^2}{N_0}}\,\rho} \qquad (4.11\text{-}3)$$

where ρ is the root mean square duration about the mean of the signal. An essentially similar result holds in general for time measurements.

$$\sigma_\tau \approx \frac{1}{\sqrt{\frac{A^2}{N_0}}\,\beta} \qquad (4.11\text{-}4)$$

where β(radians/sec) is the root mean square bandwidth about the mean frequency of the signal. Unfortunately the approximations on which this latter formula is based breakdown for a square pulse. Using a slightly modified procedure one obtains the formula[4]

$$\sigma_\tau \approx \frac{\sqrt{2}\,T}{A^2/N_0} \qquad (4.11\text{-}5)$$

which is probably a trifle optimistic, even in theory.

4.12 Linear-Sweep FM-CW Radar

The possibilities with the preceding example were rather limited. Since the bandwidth for a single pulsed-sinusoid is roughly just the reciprocal of the duration it is generally not possible to find a value of T which will simultaneously give acceptable accuracy in both range and velocity. The obvious strategy, then, is to modulate the signal so that the bandwidth can be made many times larger than the reciprocal of the duration. A simple possibility would seem to be to frequency modulate the signal, e.g., to let

$$s_0(t) = \begin{cases} \sqrt{\frac{2}{T}}\, e^{j\frac{kt^2}{2}}\,; & 0 \le t \le T \\ 0\,; & \text{elsewhere} \end{cases} \qquad (4.12\text{-}1)$$

Thus the frequency of the pulse increases linearly from ω_0 at $t = 0$ to $\omega_0 + kT = \omega_0 + W$ at $t = T$. $\psi(\tau,\omega')$ is readily computed for this waveform and the result can roughly be represented as in Fig. 5. As we can see our strategy has not been entirely successful. To be sure, we observe that we can make a measurement of τ with an accuracy the order of $4\pi/W$ (roughly an improvement of $T/2\pi$ over a pulsed-sinusoid of the same duration) but only if we know the proper value of ω. And conversely we can determine ω to an accuracy the order of $2\pi/T$ if we know the proper value of τ. But if we know neither τ nor ω the best we can determine is a relationship between τ and ω of the form $k\tau - \omega =$ constant dependent on signal received.[1] Of course this result is hardly unexpected from a physical point of view since the waveform resulting from a frequency shift and from an appropriate time delay are very similar. Compared with a pulsed-sinusoid of the same duration, what we had hoped to achieve by frequency modulation was a compression in τ' of the region in which $\psi(\tau',\omega')$ is large (black area in Fig. 3) from a width $\approx T$ to a width $\approx 4\pi/W$ without a corresponding spread in the ω' direction. What we did not anticipate perhaps is that the volume of $\psi(\tau',\omega')$, instead of being compressed as we squeezed along the τ axis, has leaked out into the first and third quadrants. Alternately what we have achieved is a rotation rather than a compression of the CW characteristic of Fig. 3.

We should not conclude, however, that FM-CW has little advantage over ordinary CW as a radar waveform. In many practical cases Ω is small compared with W and velocity information is of little use. In this case FM-CW has essentially the accuracy and freedom from ambiguities of an impulse radar having the same bandwidth, together with the lower peak power for the same energy characteristic of the CW case. Moreover, assuming that we can guarantee that we are looking at one and the same target, it may be possible to make a second measurement with a different k (e.g., $-k$) and thus to determine both τ and ω with high accuracy and without ambiguity. Nevertheless, compared with other other waveforms to be considered, the price of the FM-CW radar is high in terms of bandwidth for the results obtained. Its principle advantage is simplicity of implementation.

4.13 Coherent Periodic Pulse Radar

The commonest radar waveform, of course, is the periodic pulsed-sinusoid with or without various minor variations. We shall assume for analysis that $s_0(t)$ is real and has the form indicated in Fig. 6a. By implication we are thus assuming that carrier phase is coherent from pulse to pulse, i.e., that the pulses are merely bursts selected from the same continuous sinewave. A closely related waveform is generated physically by starting an oscillator from noise separately for each pulse so that the carrier phase is random from pulse to pulse. The performance obtained with this second, incoherent, waveform is slightly different in some respects from that to be described as will be mentioned in a later section.

$\psi(\tau,\omega)$ for the waveform of Fig. 6a is shown in Fib. 6b and 6c. Clearly the accuracy of simultaneous measurements of τ and ω is now essentially the best which can be expected for a waveform of this bandwidth and total duration. More exactly, equations (4.11-2) and (4.11-5) may be applied to this case yielding

$$\sigma_\omega = \frac{2\sqrt{3}}{T\sqrt{A^2/N_0}} \qquad (4.13\text{-}1)$$

$$\sigma_\tau = \frac{\sqrt{2}\,\delta}{A^2/N_0} = \frac{2\sqrt{2}\,\pi}{W A^2/N_0} \qquad (4.13\text{-}2)$$

---

1. Clearly the best way to describe this situation would be in terms of the parameters of an ellipse in the τ'-ω' plane.

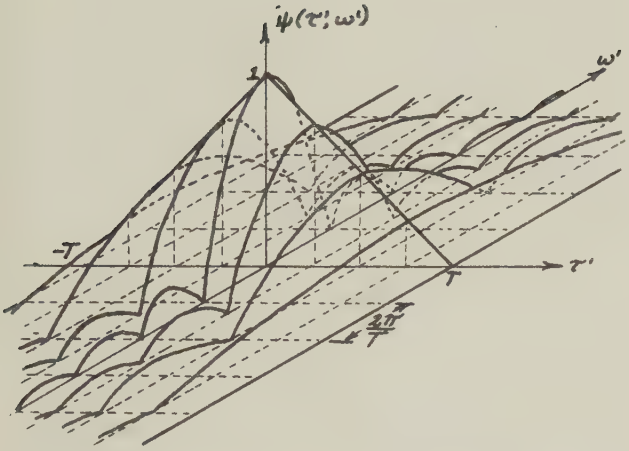


Fig. 2: $\psi(\tau';\omega')$ for Single Pulsed-Sinusoid.

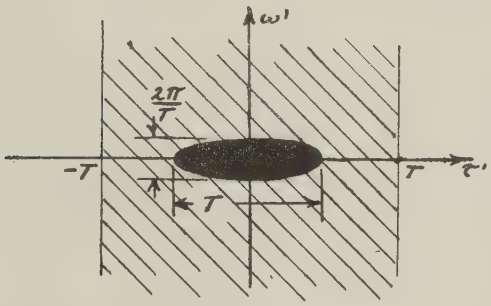


Fig. 3: $\psi(\tau';\omega')$ for CW Sinusoid.

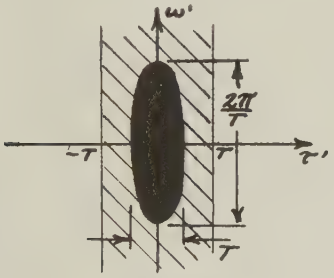


Fig. 4: $\psi(\tau';\omega')$ for Impulse of Sinusoid.

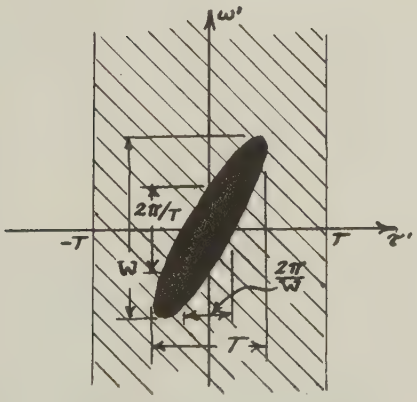
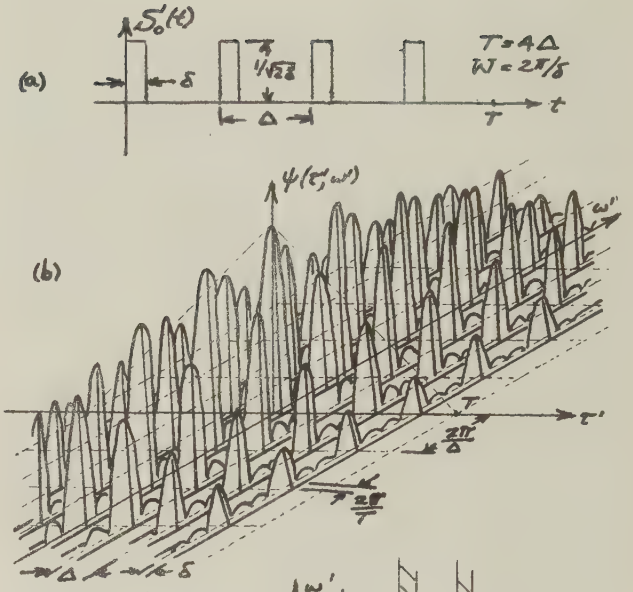


Fig. 5: $\psi(\tau';\omega')$ for Linear-Sweep FM-CW.



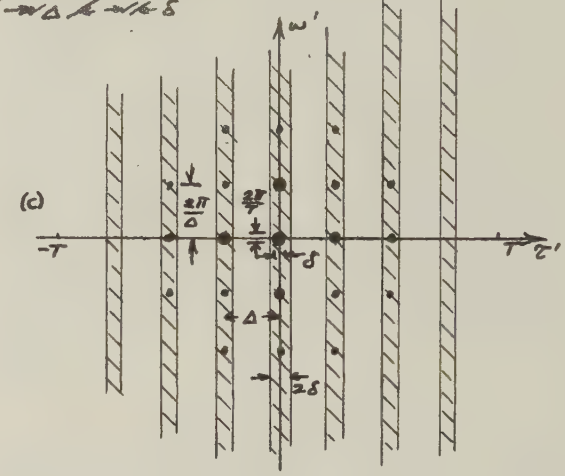


Fig. 6: $\psi(\tau';\omega')$ for Periodic Pulsed Sinusoid.

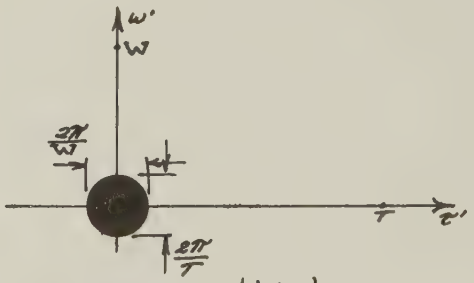


Fig. 7: "Ideal" $\psi(\tau';\omega')$ for Radar.

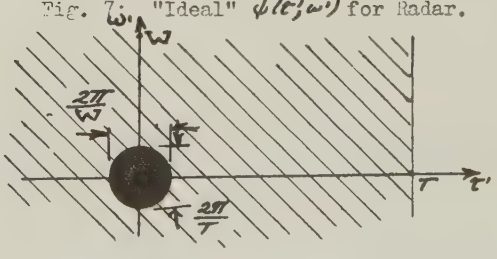


Fig. 8: $\psi(\tau';\omega')$ for Noise Waveform.

where T, δ and W have the significance indicated in Fig. 6a and A^2 is to be interpreted as the total received energy.

Although the accuracy situation, thus, leaves little to be desired in that essentially the best possible performance is obtained within the allowed bandwidth, apparently a new difficulty has appeared. Assuming that $\Theta > \Delta$ and/or $\Omega > 2\pi/\Delta$ there will now be ambiguities in the determination of τ and/or ω. These correspond to the familiar "second-time around" echoes and "blind velocities." It is true, of course, that the spikes in $\psi(\tau;\omega)$ at multiples of the repetition rate and repetition interval are smaller than the spike at the origin, and hence if the ratio A^2/N_0 is large enough, the highest spike in $P(\tau,\omega)$ should with high probability correspond to the correct parameters. Some idea as to how large A^2/N_0 must be can be acquired from the following argument. Suppose that there is only one possible alternate pair of values which might be confused with each target. Such a situation might arise, for example, if the a priori information made it possible to discard other alternatives as unlikely or impossible, e.g., $\Theta < \Delta$ and $2\pi/\delta < \Omega < 4\pi/\delta$. Let $\psi(\tau';\omega')$ be the height of the ambiguous spike. We wish, then, to compute the probability that the wrong spike in $P(\tau,\omega)$ will actually be higher (because of noise) than the correct spike. Clearly if $\psi(\tau';\omega')=1$ then there is complete equivocation; both spikes in $P(\tau,\omega)$ will under all conditions be identical in height and we can say that the probability of error is 0.5. We can compute the relationship between A^2/N_0 and $\psi(\tau';\omega')$ such that the probability of error is less than some artitrary value, say 0.10. The result [5] is approximately[1]

$$\psi(\tau';\omega') < 1 - \frac{3.4}{A^2/N_0} \qquad (4.13\text{-}3)$$

Thus if $\psi(\tau';\omega') = 0.97$, A^2/N_0 would have to be greater than ~115 before the probability of error would be less than 10%. Of course (4.13-3) is a theoretical result and ignores such questions as distortion and drift. Independent of the value of A^2/N_0, it would be necessary to go to a great deal of trouble to build practical equipment capable of distinguishing reliably and over a wide dynamic range a difference as small as 3%. For most practical purposes spikes in $\psi(\tau;\omega)$ greater than, say, 0.5, constitute unresolvable ambiguities.

Finally the presence of ambiguities has a significant effect on the value of M to use in equation (3.2-2) for estimating the reliability of detection. If both $\Theta < \Delta$ and $\Omega < 2\pi/\delta$ then it is easy to argue that the appropriate value of M is roughly

$$M \approx \frac{\Theta}{\delta} \cdot \frac{\Omega T}{2\pi} \qquad (4.13\text{-}4)$$

On the other hand if $\Theta \geq \Delta$ and $\Omega \geq \frac{2\pi}{\delta}$ then not all of the signals given by (4.13-4) will be independent so that the appropriate value of M is

$$M \approx \frac{2\pi/\Delta}{2\pi/T} \cdot \frac{\Delta}{\delta} = T \cdot \frac{1}{\delta} \qquad (4.13\text{-}5)$$

i.e., equal to the time-bandwidth product for the signal. Considering the implications of the Sampling Theorem [11] concerning the number of degrees of freedom in a signal of limited time and frequency duration, this is an entirely reasonable result.

4.14 The Radar Uncertainty Principle [1]

It should at this point be obvious that $\psi(\tau';\omega')$ corresponding to the ideal radar waveform should have the appearance of Fig. 7--a single narrow spike at the origin and nothing anywhere else in the plane. For maximum accuracy the spike should have a width of approximately $2\pi/W$ in frequency and $2\pi/W$ in time, and T and W should be independently adjustable. The difficulty is--as we might expect--that such conditions are fundamentally impossible to achieve. We now wish to study why.

Our efforts thus far to achieve a waveform having a $\psi(\tau;\omega)$ similar to Fig. 7 have been kind of like squeezing a pillow--as we push in one direction the pillow bulges out in the other, and if we are too persistent the casing breaks and we have piles of feathers all over the landscape. Or perhaps a better analogy would be to imagine that the $\psi(\tau;\omega)$ contour is the surface of a pile of sand. As we adjust the waveform we seem to be able to move the sand around but unable to get rid of any of it. This latter analogy, with one modification--namely that we should talk about the $\psi^2(\tau;\omega)$ contour instead of the $\psi(\tau;\omega)$ contour, is actually a precise statement of the most important limitation on radar accuracy-ambiguity performance, i.e., what we shall call the Radar Uncertainty Principle. But before we give a precise formulation of this principle it is perhaps valuable to demonstrate it in an approximately quantitativemanner for the various waveforms we have thus far considered. For most radar waveforms $\psi(\tau;\omega)$ consists roughly of a number of spikes of approximately unit height together with regions in which $\psi(\tau;\omega) \approx 0$. An approximate evaluation of the volume under the $\psi^2(\tau;\omega)$ contour can be achieved by replacing the spikes with roughly equivalent cylinders of unit height and ignoring the volume in the regions where $\psi(\tau;\omega) \approx 0$. For the three waveforms thus far considered this crude volume computation is as follows:

	[Base Area Of Unit Height Cylinders]	×	[Approximate No of Cylinders]	= Volume
CW:	$(T) \times \left(\frac{2\pi}{T}\right)$	×	1	$= 2\pi$
PRF-CW:	$\left(\frac{2\pi}{W}\right) \times \left(\frac{2\pi}{T}\right)$	×	$\frac{T}{2\pi/W}$	$= 2\pi$
Pulse:	$\delta \times \left(\frac{2\pi}{T}\right)$	×	$\left(\frac{T}{\Delta}\right) \times \frac{(2\pi/\delta)}{(2\pi/\Delta)}$	$= 2\pi$

Although it is something of an accident that this rather crude method works so nicely in these cases,

1. Actually (4.13-3) applies to the case of known initial phase angle, but the error in A^2/N_0 resulting from applying (4.13-3) to the random initial phase case is small.

the conclusions nevertheless are correct. More formally, a precise statement of the Radar Uncertainty Principle is:

Independent of the form of $\psi(\tau';\omega')$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}d\omega'\int_{-\infty}^{\infty}\psi^2(\tau';\omega')\,d\tau' = 1 \qquad (4.14\text{-}1)$$

This result is easily proved by directly carrying out the indicated integrations after substituting the definition of $\psi(\tau';\omega')$ from (4.1-1).

The Radar Uncertainty Principle has a number of important applications:

1. As a part of the a posteriori probability approach, the Uncertainty Principle helps to emphasize that waveform selection, rather than ingenuity in detector design, is the determining element in radar accuracy, ambiguity, and (as we shall see) resolution performance.
2. By setting a bound on performance quality, it prevents much fruitless searching for waveform and detection methods intended to achieve such impossible performance as that described at the beginning of this section.
3. The Radar Uncertainty Principle has proved very helpful in finding the flaws in various suggested radar waveforms, detection procedures, MTI schemes, etc. Specifically, one can be sure that the analysis is complete if and only if all of the volume under the $\psi^2(\tau';\omega')$ contour has been accounted for.

Unfortunately, although the Radar Uncertainty Principle represents a necessary condition for the existence of a waveform $s_0(t)$ having a given $\psi(\tau';\omega')$ it is not sufficient. A number of additional conditions can be specified, including sufficient conditions, but the forms of these conditions are not sufficiently simple to be really useful in waveform synthesis. There seems to be no real substitute at this point for an educated but intuitive guess followed by careful analysis.

4.15 The Ideal Waveform

We shall conclude this section with a discussion of several waveforms which come as close as possible to the ideal radar waveform--at least from the standpoint of accuracy and freedom from ambiguities. These waveforms have essentially the $\psi(\tau';\omega')$ plot of Fig. 8, i.e., a single central spike of width $2\pi/T$ in frequency and $2\pi/W$ in time (where T and W are the echo duration and bandwidth in rad/sec respectively) with the remainder of the necessary volume (if $TW/2\pi$ is large this will be nearly all the volume) spread out more-or-less uniformly over a region roughly T wide in time and W wide in frequency.

All of the waveforms corresponding to Fig. 8 have in common that they are in some sense noisy or pseudo-random--by which we mean that it takes many numbers to specify them as opposed to the waveforms we have already considered which are specified by just a few numbers, e.g., pulse length, repetition rate, etc. For example, if $s_0(t)$ is a sample from almost any sort of noise, e.g., hard-limited narrow-band Gaussian noise of duration T and bandwidth W, the corresponding $\psi(\tau';\omega')$ will with high probability look like Fig. 8 provided that $TW/2\pi$ is very large, the order of 10^4 to 10^6 or so. But when $TW/2\pi$ is only the order of 10^2 to 10^3, the noise waveform has to be selected with some care if spikes of height 0.5 or more at values of τ' and ω' other than the origin are to be avoided.

An interesting example of a suitable waveform having $TW/2\pi = 10^3$ or less is the coded-pulse waveform.[5] This waveform is constructed by starting with a pulsed sinusoid of duration T. This pulse is then divided into $TW/2\pi$ intervals, each of duration $2\pi/W$. Each interval is then preserved as originally, or reversed in phase by 180°, according to whether the corresponding position in a binary sequence or code of length $TW/2\pi$ is 0 or 1. Clearly the performance of such a waveform then depends on the code selected. Almost any phase-modulated waveform having a bandwidth W or less can be closely approximated by choosing the proper code. At the moment we are interested in that code or codes which for a given value of $TW/2\pi$ will minimize $\psi(\tau',\omega')$ for $\tau' \neq 0, \omega' \neq 0$. The problem of finding such a code is closely related to the coding problem in information theory and many of the methods employed there are applicable. For example, for $TW/2\pi = 2^n - 1$ the best codes yet found (and there is reason to believe no better codes exist) are those called by various authors [6][7] shift-register or null sequence codes of maximum length. An example of such a code for n = 5, $TW/2\pi = 31$ is the following

0111010000110010011111011100001 01

which is obtained by starting off with the code of length 5, 01101 (any other starting code except 00000 will yield equivalent results), and setting the next element equal to the sum modulo 2 of the first, second, third, and fifth digits preceding. This process is repeated for each successive element. It will be found that after 31 elements have been written down, the sequence will repeat, and indeed, the fact that the code does not repeat prior to the 2^n-1 element is a sufficient test that a proper parity check rule has been employed. For any value of n there are a number of such codes--all of which, so far has been determined, are equally satisfactory for our present purposes. It should be recalled that we are interested, loosely, in minimizing the maximum value of $\psi(\tau';\omega')$ for all values of τ' and ω' excluding the origin. In particular it would not be sufficient alone to minimize $\psi(\tau';\omega')$ along the τ' axis, i.e., for $\omega' = 0$. Codes can be found which have better performance along the τ' axis than the maximum length null sequences, but such codes inevitably seem to have large spikes, i.e., potential ambiguities, off the τ' axis. Experience would seem to suggest that maximum-length null sequence codes yield a maximum value of $\psi(\tau';\omega')$ (excluding the origin) the order of $\sqrt{2\pi/TW}$.

We are justified in concluding that--from

standpoint of accuracy and ambiguity--coded-pulse or other noise-like waveforms achieve essentially the ultimate possible performance. Why then have such waveforms not had a wider application in radar? To be sure, an appreciation of the value of such waveforms is relatively recent and there is a feeling, which we do not completely share, that the equipment to generate and process such waveforms is impractically complicated. But the most important reason is that noise waveforms have in many practical target situations serious disadvantages from the standpoint of resolution. This is a problem which we now wish to investigate in general.

5.0 Resolution

Thus far we have considered only cases in which it was known a priori that at most one target echo was present at any one time. Such a fortuitous situation almost never occurs in practice. At the very least we have to discriminate against our own transmitted signal which represents a huge signal at zero range and velocity. Moreover, in many cases ground clutter, chaff, the ionosphere, meteor trails, etc.. return echoes which not only contain little useful information but which, if they have a large amplitude, may obscure the echoes from desired targets. Indeed, it is probably only a slight exaggeration to claim that in many cases the problem of resolving desired echoes from undesired echoes and from one another is so important as to be the principal requirement on the radar design. Freedom from ambiguities, for example, might be a nice thing to have, but not if it can be obtained only with a reduction in resolution performance.

Despite the importance of this problem, there is remarkably little we can say about it of any general validity of utility. There are, of course, several obvious platitudes which despite their triviality help to formulate the problem.

1. If the interfering signal is known complete and exactly (i.e., if A, τ, ω, and the carrier phase angle are known precisely) then there is really no resolution problem since the obvious and theoretically correct procedure is to subtract a replica of the interfering signal from $r(t)$ prior to processing.

2. If the class of undesired signals is identical with the class of desired signals then there is obviously no possibility for resolution.

Thus the resolution problem is theoretically and practically an interesting one only if the exact characteristics of the interfering signals are unknown in one or more respects and different in one or more respects from the desired signals. There are, of course, many kinds of situations, but a large fraction of these are of little interest. For example, two signals, which are identical except for amplitude obviously can be resolved with high reliability only if the most probable desired signal energy is much greater than the most probable undesired signal energy. The most interesting cases are those in which the signals to be distinguished differ in time delay and/or frequency shift. Here the appropriate measure of the difference between signals is $\psi(\tau,\omega)$ and there are two cases of interest:

1. Each desired signal is orthogonal to the entire class of undesired signals;
2. Each desired signal is at most weakly correlated with some of the members of the class of undesired signals.

Clearly, resolution is almost impossible if the undesired signals are large and strongly correlated with the desired signals.

In the first case resolution is trivially simple to obtain. It can easily be shown that the a posterior probability that any particular desired signal is present is entirely independent of the presence or size of the undesired signals. Hence all of the analysis in preceding sections with respect to reliability of detection, accuracy, and ambiguity is immediately applicable. Resolution in this case is obtained automatically.

On the other hand, in the second case the situation is not nearly so clear. We might, of course, just pretend that the signals to be separated are orthogonal and thus build our decision circuits as previously. The resolution performance under these conditions will be quite good, i.e., P_d and P_f for the desired signals will be essentially unaltered by the presence of the undesired signals, provided that the ratio of the energy of the undesired signal to the energy of the desired signal remains somewhat less than $1/\psi(\tau,\omega)$. Thus a degree of relative resolution[1] can be obtained. However, it should be possible to achieve somewhat better relative resolution by altering the form of the detector, e.g., intentional using an appropriately mis-matched filter. Of course, the reliability of detection for desired signals will then be less, but in the presence of correlated undesired signals such a loss in reliability of detection is inevitable. In certain hypothetical cases the best possible form for the mis-matched filter can be worked out. For example, if the class of undesired signals consists of a finite set of signals at known discrete values of τ and ω but with unknown amplitude it is possible to design a detector providing essentially infinite resolution at only a (usually) small price in detectability. But for the case of continuous parameter distributions no really satisfactory procedures are known. However it seems unlikely that any truly significant improvements in performance can be achieved by mismatching the filter. Emperical methods for designing clutter rejection filters, for example, are probably as good as any.

The conclusion is inescapable that if resolution is important the radar waveform must be so chosen as to make the signals to be resolved as nearly orthogonal as possible. From this point of view the periodic pulse waveform has outstanding advantages constituting as good a reason as any for its overwhelming popularity. By permitting ambiguities, the periodic pulse waveform manages to cram nearly all the volumes required under the

1. Relative resolution is clearly what we have in mind when we speak, for example, of sub-clutter visibility.

$\phi(\tau,\nu)$ surface into tall slender spikes, leaving most of the τ-ν plane absolutely empty. No other waveform is quite so well suited for those applications (e.g., GCI radars) in which resolution capability is (or should be) the pre-eminent design specification. However, when slight compromises can be tolerated in resolution performance, important advantages in other respects can be achieved by several variations in the coherent periodic pulse waveform, e.g., non-coherent phase from pulse-to-pulse, a periodically-repeated phase-or-frequency modulated pulse, or a time-duplexed radar employing two repetition rates, each for half the echo duration. Other schemes, e.g., staggered repetition rate, or changing frequency from pulse-to-pulse, are less useful since they have a serious effect on resolution performance.

We do not intend to create the impression that a periodic pulse radar is an adequate solution to the usual, let alone the extreme, resolution situation; it is not. There are many radar systems in particular which fail to achieve the desired performance, if for no other reason, because either the resolution performance of the periodic pulse radar is inadequate or because the ambiguities associated with this waveform are intolerable. We feel rather strongly that within the constraints imposed by our model, i.e., by our present radar system philosophy, a satisfactory solution to these problems is essentially impossible. In particular we believe that it will be necessary to break away from the idea of a finite (i.e., short) observation interval with its associated concept of the "occupied cell" in time and frequency. Our desired targets have in general a time history or life pattern which is a much more distinguishing characteristic than their instantaneous position and velocity. We shall have to learn how to exploit this characteristic The philosophical, computational, and practical problems appear, at this time, to be rather difficult, but here, if anywhere, would seem to lie the future promise of radar.

6.0 Acknowledgements

Although the author takes full responsibility for the opinions and results presented in this paper, little if any of the paper can truly be said to be original. In particular the now classic work of P. M. Woodward and I. L. Davies underlies the entire paper, both in basic approach and, in some cases, in detail. In addition the author has profited immeasurably from numerous papers, reports, and discussions with many people --more than he can possibly acknowledge or perhaps in some cases even recall. Specifically the author wishes to acknowledge the cooperation and contributions of his colleagues at Lincoln Laboratory and M.I.T., notably Prof. R. M. Fano, R. M. Lerner, L. G. Kraft, R. Manasse, and F. A. Rodgers, among many others.

7.0 Bibliography

The following list contains only those books and papers referred to in the text or deemed most appropriate to the matters under discussion. There are numerous other excellent and important papers on various aspects of this subject, most of which are listed in the bibliographies of the references cited.

[1] P. M. Woodard, "Probability and Information Theory, with Applications to Radar", McGraw-Hill Book Co., New York, N.Y., 1955.
P. M. Woodard and I. L. Davies, "A Theory of Radar Information", Phil. Mag., vol.41, pp. 1001-1017, Oct., 1950.
P. M. Woodard and I. L. Davies, "Information Theory and Inverse Probability in Telecommunication", Proc. I.E.E., Pt. III, vol. 99, pp. 37-44, Mar., 1952.

[2] W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The Theory of Signal Detectability", Trans. PGIT-4, I.R.E., Sept., 1954.

[3] S. O. Rice, "Statistical Properties of a Sine Wave Plus Random Noise", B. S. T. J., vol. 27, pp. 109-157, Jan., 1948.

[4] R. Manasse, "Range and Velocity Accuracy from Radar Measurements", unpublished internal report, Lincoln Laboratory, Mass. Inst. of Tech., Cambridge, Mass., Feb., 1955.

[5] C. W. Helstrom, "The Resolution of Signals in White Gaussian Noise", Proc. I.R.E., vol. 43, pp. 1111-1118, Sept, 1955.

[6] D. A. Huffman, "The Synthesis of Linear Sequential Coding Networks", Proc. Third London Symposium on Information Theory, Sept., 1955.

[7] N. Zierler, "Several Binary-Sequence Generators", Tech. Rep. 95, Lincoln Laboratory, Mass. Inst. of Tech., Cambridge, Mass., Sept., 1955.

INFORMATION FOR AUTHORS

Authors are requested to submit editorial correspondence or technical manuscripts to the Publications Chairman for possible publication in the PGIT TRANSACTIONS. Papers submitted should include a statement as to whether the material has been copyrighted, previously published, or accepted for publication elsewhere.

Papers should be written concisely, keeping to a minimum all introductory and historical material. It is seldom necessary to reproduce in their entirety previously published derivations, where a statement of results, with adequate references, will suffice.

To expedite reviewing procedures, it is requested that authors submit the original and two legible copies of all written and illustrative material. The manuscript should be double-spaced, and the illustrations drawn in india ink on drawing paper or drafting cloth. Each paper should include a carefully written abstract of not more than 200 words. Upon acceptance, papers should be prepared for publication in a manner similar to those intended for the PROCEEDINGS OF THE IRE. Further instructions may be obtained from the Publications Chairman. Material not accepted for publication will be returned.

IRE TRANSACTIONS ON INFORMATION THEORY is published four times a year, in March, June, September, and December. A minimum of one month must be allowed for review and correction of all accepted manuscripts. A period of approximately two months additional is required for the mechanical phases of publication and printing. Therefore, all manuscripts must be submitted three months prior to the respective publication dates. In addition, the IRE CONVENTION RECORD is published in July, and a bound collection of Information Theory papers delivered at the annual IRE National Convention is mailed gratis to all PGIT members.

All technical manuscripts and editorial correspondence should be addressed to Laurin G. Fischer, Federal Telecommunications Lab., 492 River Road, Nutley, N. J. Local Chapter activities and announcements, as well as other nontechnical news items, should be addressed to Nathan Marchand, Marchand Electronic Labs., Riversville Road, Greenwich, Conn.